Task-oriented and Semantic-aware Communications and Networking for 6G

Task-oriented and Semantic-aware Communication for Edge Video Analytics

Jun Zhang

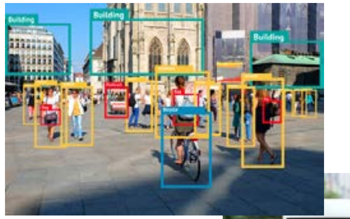
香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

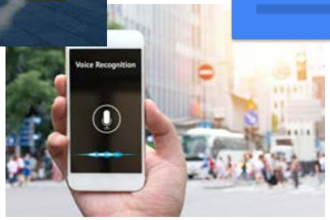
Outline

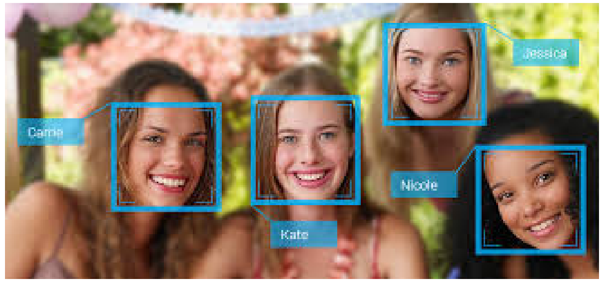
- Background: Edge inference and task-oriented communication
- Case studies
  - Task-oriented communication for edge-assisted inference via information bottleneck (IB)
  - Task-oriented communication for cooperative perception via distributed information bottleneck (DIB)
  - Task-oriented communication for edge video analytics (sequential data)
- Conclusions

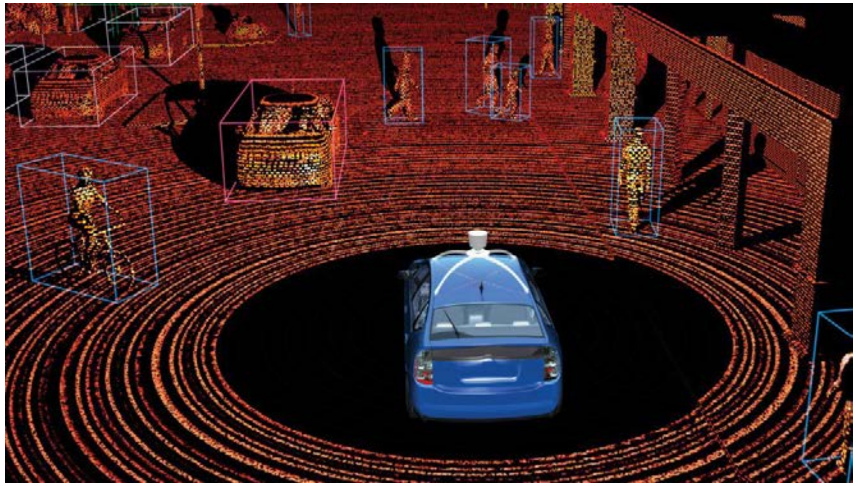
Edge inference and task-oriented communication

Edge inference



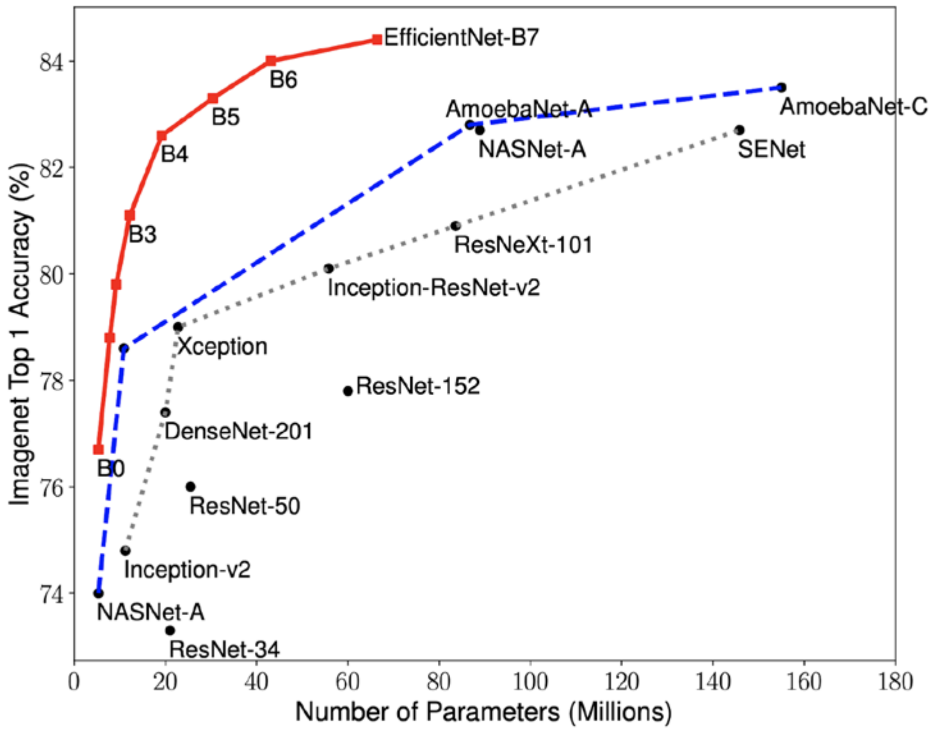




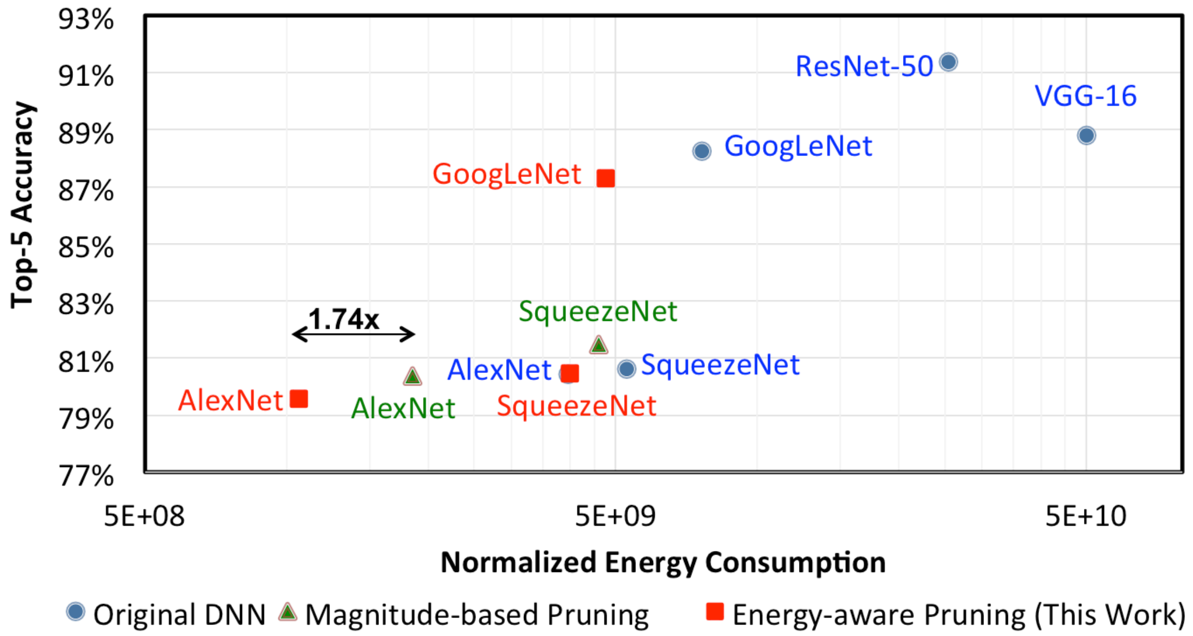


Challenges of edge inference

LARGE size of DNN models



HIGH energy of DNN models



https://ai.googleblog.com/2019/05/efficientnet-improving-accuracy-and.html, May 2019

Challenges of edge inference

A single device is limited in

- onboard computing resources;
- limited perception capability;
- limited energy supply.

Effective communication is critical to

- access external computing power;
- improve perception capability;
- prolong battery time;
- overcome partial observation.

Edge AI

amazon web services
AWS IoT for the Edge
Bring cloud connectivity and intelligence to the IoT edge

Microsoft Azure
Azure IoT Edge
Cloud intelligence deployed locally on IoT edge devices

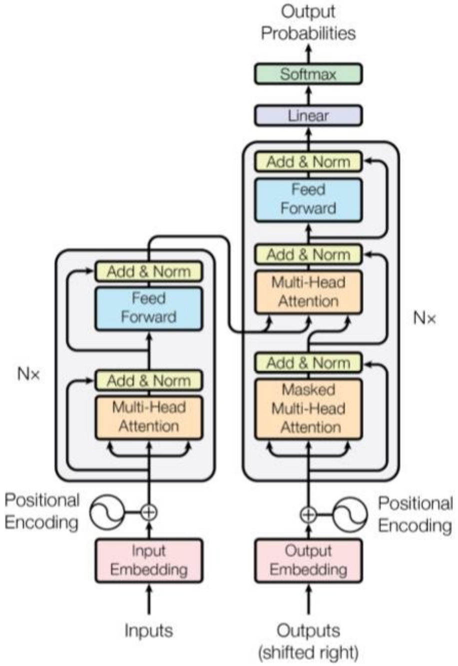
Google Cloud
Cloud IoT Edge ALPHA
Deliver Google AI capabilities at the edge.

Computing

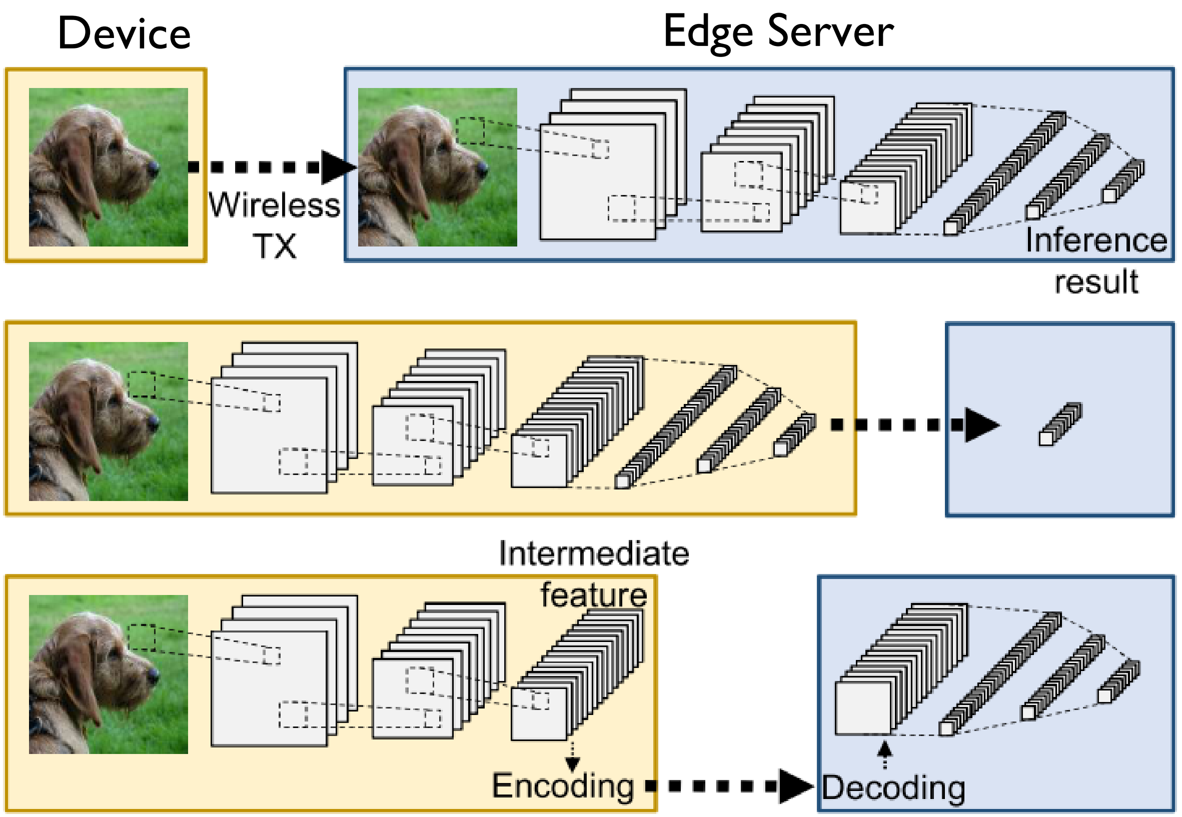
Communication

Deep Learning

Edge AI



Solutions for edge inference



Server-based method

- ☒ High communication load
- ☒ Privacy concern

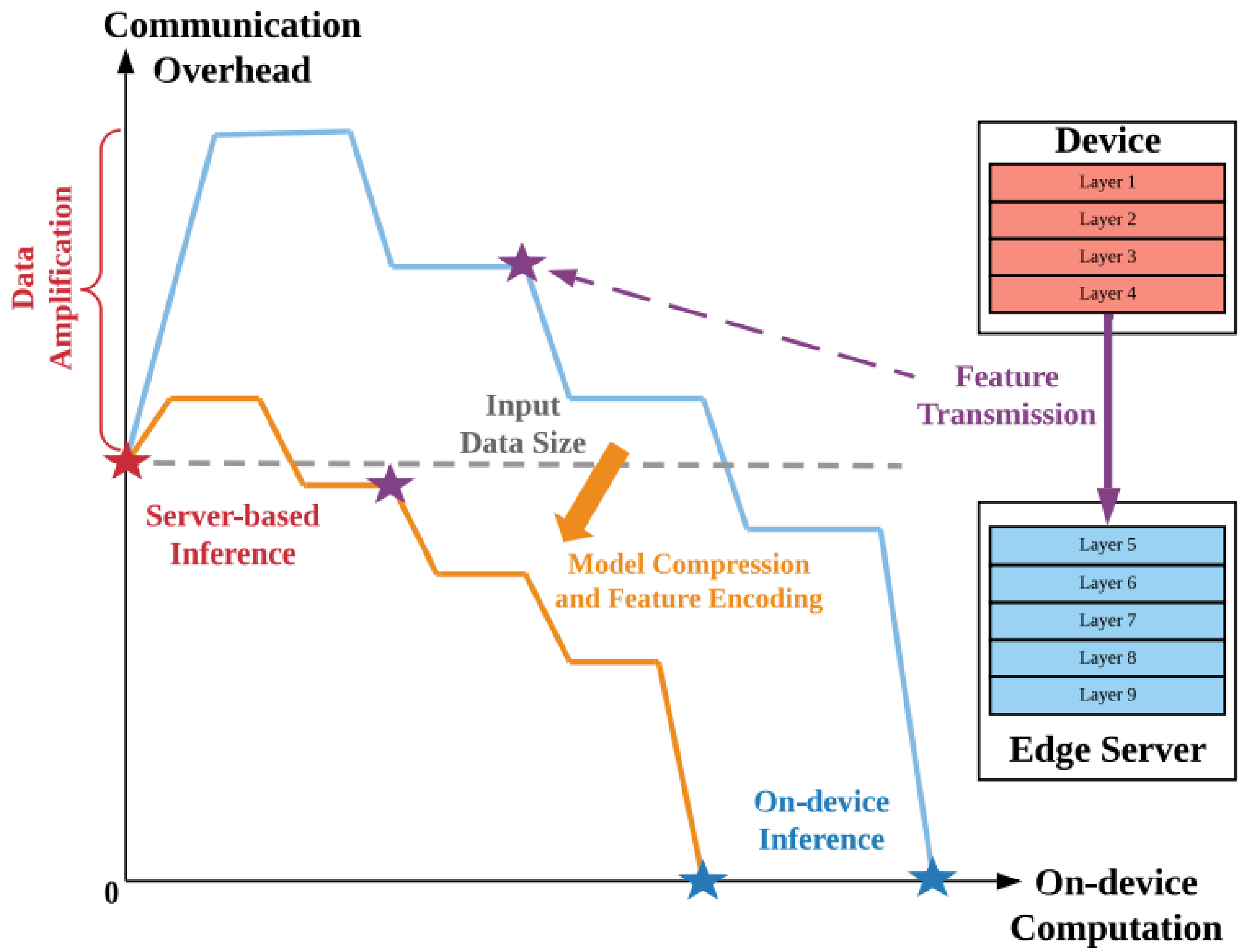
On-device processing

- ☒ High local computation
- ☒ Limited performance

Device-edge co-inference

Balance communication and local computation

Device-edge co-inference with model partitioning



Feature encoding is critical for communication-computation tradeoff

New communication problem

- Communication for edge inference (not for data reconstruction)

Three levels of communications

Shannon's information theory

Level A
The technical problem

- How *accurately* can the symbols of communication be transmitted?

How to communicate?

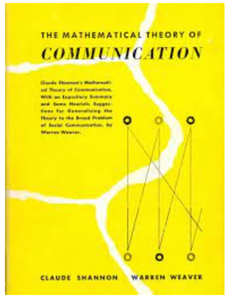
Level B
The semantic problem

- How *precisely* do the transmitted symbols convey the desired meaning?

What to communicate?

Level C
The effectiveness problem

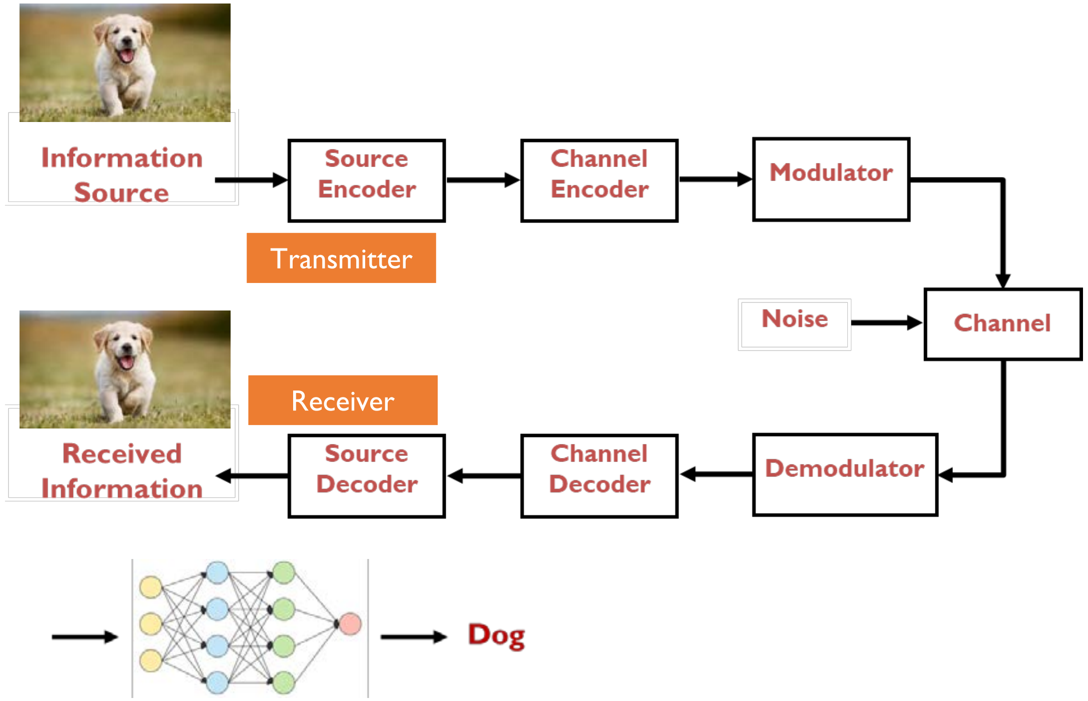
- How *effectively* does the received meaning affect conduct in the desired way?



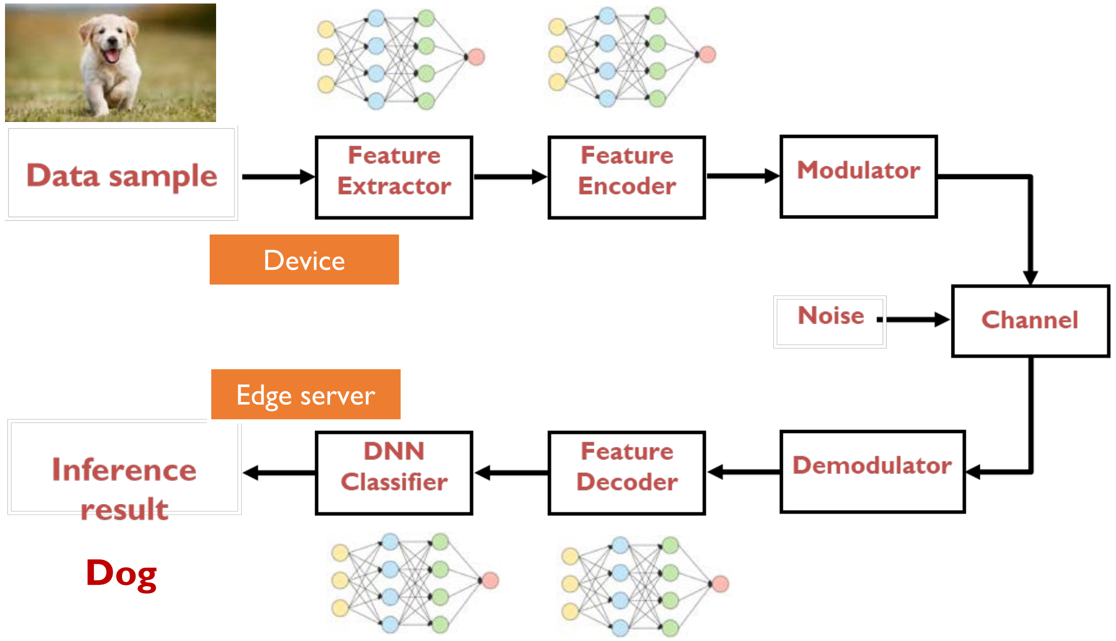
W. Weaver. Recent contributions to the mathematical theory of communication. In C. E. Shannon and W. Weaver, editors, *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, 1949.

Data-oriented vs. Task-oriented communication

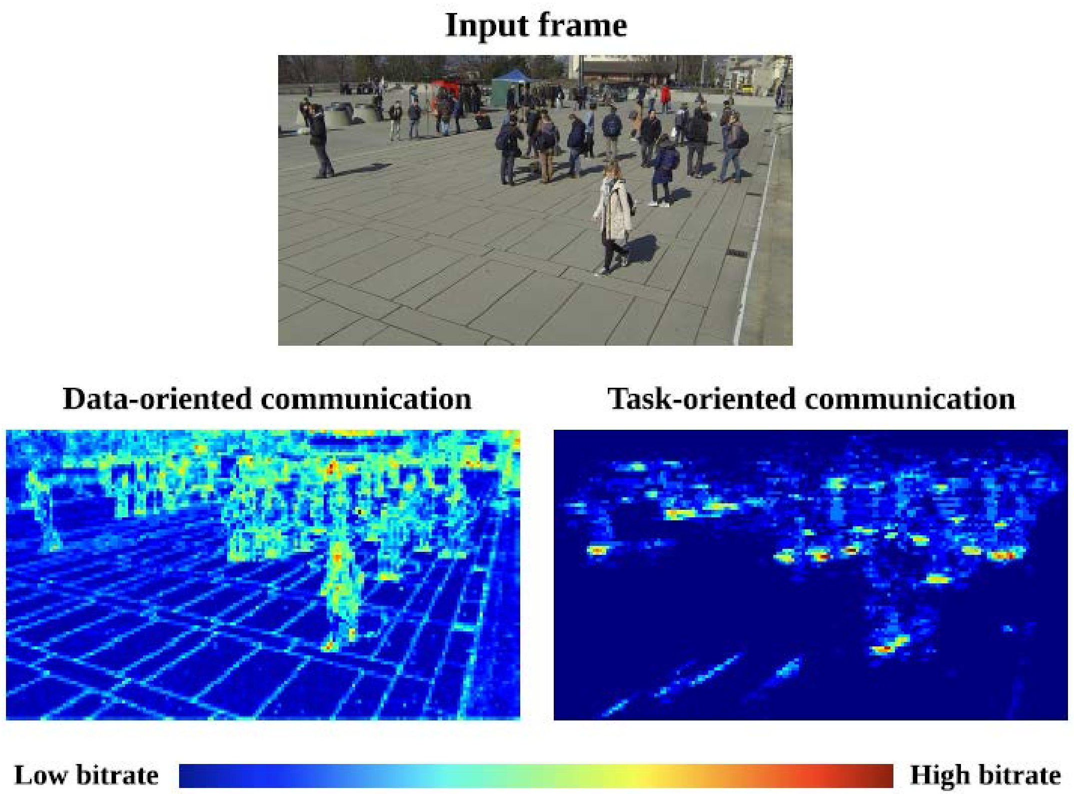
Data-oriented Communication



Task-oriented Communication



Example: Multi-camera pedestrian occupancy prediction



Data-oriented communication:
- It allocates many bits to represent the background and ground texture.
- However, these details almost do not influence the performance of the downstream task.

Task-oriented communication:
- It focuses on task-relevant information (e.g., the foot points of pedestrians) and discards the redundancy.
- It substantially reduces the communication overhead and latency.

Task-oriented communication system design

- Design goal: To transmit *concise* and *informative* feature with *low-complexity* encoder for *low-latency high-accuracy* inference

- Theoretical foundation: source coding theory

Design challenges
- Unknown high-dimensional data distribution
- Intractable task-specific distortion metric
- High computational complexity

Design tools
- End-to-end deep learning
- Variational approximation (to make the objective tractable)
- Neural network architecture optimization

Task-oriented communication for edge inference via information bottleneck

J. Shao, Y. Mao, and **J. Zhang**, "Learning task-oriented communication for edge inference: An information bottleneck approach," *IEEE J. Select. Areas Commun.*, vol. 40, no. 1, pp. 197-211, Jan. 2022.

Naftali Tishby

(1952 –August 2021)

The information bottleneck (IB) problem

Y **Target** → X **Data** → T **Representation**

$$\min -I(Y;T) + \beta I(X;T)$$

How well T predicts Y

To promote accuracy

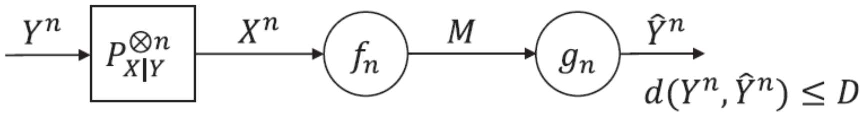
How much T compresses X

To promote generalization

Tradeoff

Preserving "relevant" information vs. finding "compact" representation

- Closely related to **remote source coding.**

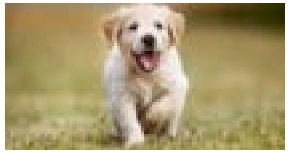


- Applications of information bottleneck
  - IB theory for deep learning
  - IB as optimization objective (to improve generalization, robustness)

N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," Annu. Allerton Conf. Commun. Control Comput., 1999.

Task-oriented communication vs. Information bottleneck

Task-oriented Commun.

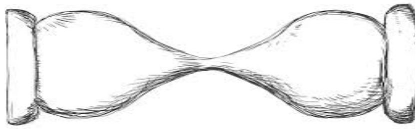


$$X \xrightarrow{\text{Encode}} Z \xrightarrow{\text{Transmission}} \hat{Z} \xrightarrow{\text{Decode}} Y$$

Input data | Transmitted feature | Received feature | Target variable

Dog

Information Bottleneck



X → T → Y

Data | **Feature** | **Target**

Task-oriented communication via the IB principle

$$\underbrace{Y}_{\text{Target}} \leftrightarrow \underbrace{X}_{\text{Input}} \xrightarrow{\text{Encoder}} \underbrace{Z}_{\text{Feature}} \xrightarrow{\text{Channel}} \underbrace{\hat{Z}}_{\text{Rec. Feature}} \xrightarrow{\text{Decoder}} \underbrace{\hat{Y}}_{\text{Ouput}}$$

$$\min \underbrace{-I(\hat{Z}, Y)}_{\text{Distortion}} + \beta \cdot \underbrace{I(\hat{Z}, X)}_{\text{Rate}}$$

- We do not need to recover X from $\hat{Z}$
- $\hat{Z}$ only needs to retain task-relevant information to infer Y

How well $\hat{Z}$ predicts Y

To promote accuracy

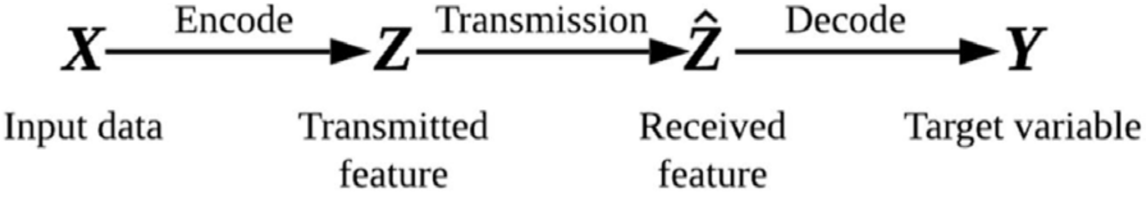
How much $\hat{Z}$ compresses X

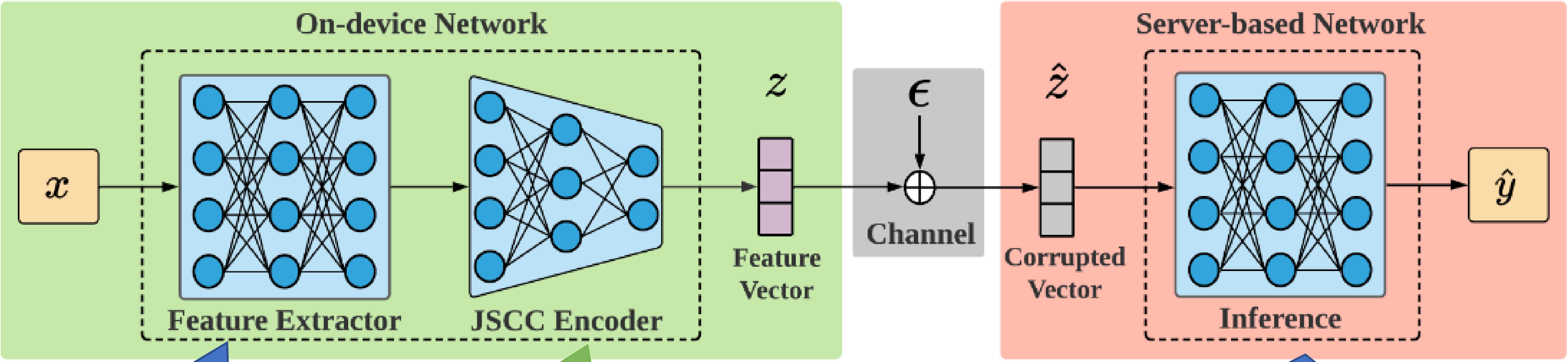
To reduce communication overhead

Relevance-rate tradeoff

- Main design challenges:
  - How to estimate mutual information?
  - How to effectively control communication overhead?
  - How to handle dynamic channel conditions?

Variational Feature Encoding (VFE)





Lightweight feature extractor: to control on-device computation/energy

Joint-source-channel coding (JSCC) encoder: design component, to minimize the output dimension

Powerful server-side network

VFE: Variational approximation

$$\underbrace{Y}_{\text{Target}} \leftrightarrow \underbrace{X}_{\text{Input}} \rightarrow \underbrace{Z}_{\text{Feature}} \rightarrow \underbrace{\hat{Z}}_{\text{Rec. Feature}} \rightarrow \underbrace{\hat{Y}}_{\text{Ouput}}$$

Intractable objective

$$\boxed{-I(Y,\hat{Z}) + \beta I(\hat{Z},X)} = -\int p(\boldsymbol{y} \mid \hat{\boldsymbol{z}}) p(\hat{\boldsymbol{z}}) \log p(\boldsymbol{y} \mid \hat{\boldsymbol{z}}) d\boldsymbol{y} d\hat{\boldsymbol{z}} + \beta \int p_\phi(\hat{\boldsymbol{z}} \mid \boldsymbol{x}) p(\boldsymbol{x}) \log \frac{p_\phi(\hat{\boldsymbol{z}} \mid \boldsymbol{x})}{p(\hat{\boldsymbol{z}})} d\boldsymbol{x} d\hat{\boldsymbol{z}} - H(Y)$$

Variational bound

$$\leq \underbrace{-\int p(\boldsymbol{y} \mid \hat{\boldsymbol{z}}) p(\hat{\boldsymbol{z}}) \log q_\theta(\boldsymbol{y} \mid \hat{\boldsymbol{z}}) d\boldsymbol{y} d\hat{\boldsymbol{z}}}_{\text{Cross-Entropy}} + \underbrace{\beta \int p_\phi(\hat{\boldsymbol{z}} \mid \boldsymbol{x}) p(\boldsymbol{x}) \log \frac{p_\phi(\hat{\boldsymbol{z}} \mid \boldsymbol{x})}{q(\hat{\boldsymbol{z}})} d\boldsymbol{x} d\hat{\boldsymbol{z}}}_{\text{KL-Divergence}} - \underbrace{H(Y)}_{\text{constant}}$$

Variational Information Bottleneck (VIB) objective

$$\mathcal{L}_{VIB}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \mathbf{E}_{p(\boldsymbol{x},\boldsymbol{y})} \left\{ \mathbf{E}_{p_{\boldsymbol{\phi}}(\hat{\boldsymbol{z}}|\boldsymbol{x})} \left[-\log q_{\boldsymbol{\theta}}(\boldsymbol{y}|\hat{\boldsymbol{z}}) \right] + \beta D_{KL} \left(p_{\boldsymbol{\phi}}(\hat{\boldsymbol{z}}|\boldsymbol{x}) \| q(\hat{\boldsymbol{z}}) \right) \right\}.$$

Empirical estimation

$$\simeq \frac{1}{M} \sum_{m=1}^{M} \left\{ \frac{1}{L} \sum_{l=1}^{L} \left[-\log q_{\boldsymbol{\theta}} \left(\boldsymbol{y}_{\boldsymbol{m}} | \hat{\boldsymbol{z}}_{\boldsymbol{m},\boldsymbol{l}} \right) \right] + \beta D_{KL} \left(p_{\boldsymbol{\phi}} \left(\hat{\boldsymbol{z}} | \boldsymbol{x}_{\boldsymbol{m}} \right) \| q(\hat{\boldsymbol{z}}) \right) \right\}$$

- Variational approximations
  - $p_\phi(\hat{z}|x)$ is defined by the neural network (encoder)
  - $q_\theta(y|\hat{z})$ is a variational distribution to approximate $p(y|\hat{z})$
  - $q(\hat{z})$ is a variational distribution to approximate $p(\hat{z})$

Some details: Approximated closed-form solution

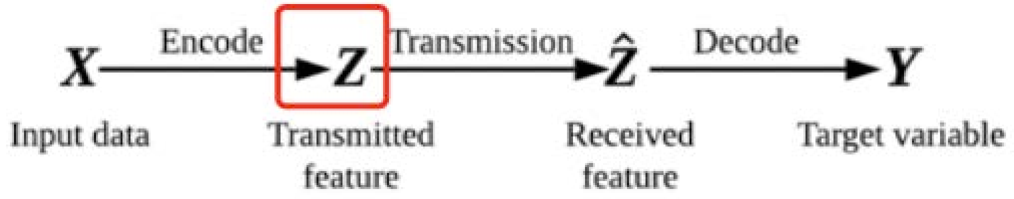
$p_\phi(\hat{z}|x)$ is a factorized Gaussian distribution

$q(\hat{\boldsymbol{z}})$ is a log-uniform distribution

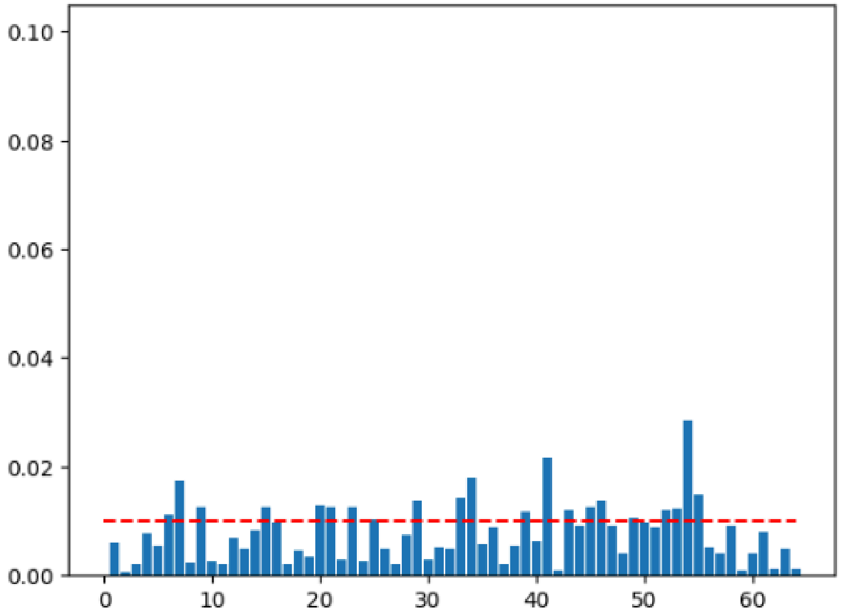
$$D_{KL}(p_\phi(\hat{z}|x)\|q(x)) = \sum_{i=1}^{n} D_{KL}(p_\phi(\hat{z}_i|x)\|q(\hat{z}_i))$$

$$\begin{aligned} -D_{KL}(p_\phi(\hat{z}_i \mid \boldsymbol{x})\|q(\hat{z}_i)) &= \frac{1}{2}\log\alpha_i - \mathbb{E}_{\epsilon\sim\mathcal{N}(1,\alpha_i)}\log|\epsilon| + \mathrm{C} \\ &\approx k_1 S(k_2 + k_3\log\alpha_i) - 0.5\log\left(1+\alpha_i^{-1}\right) + \mathrm{C} \end{aligned}$$

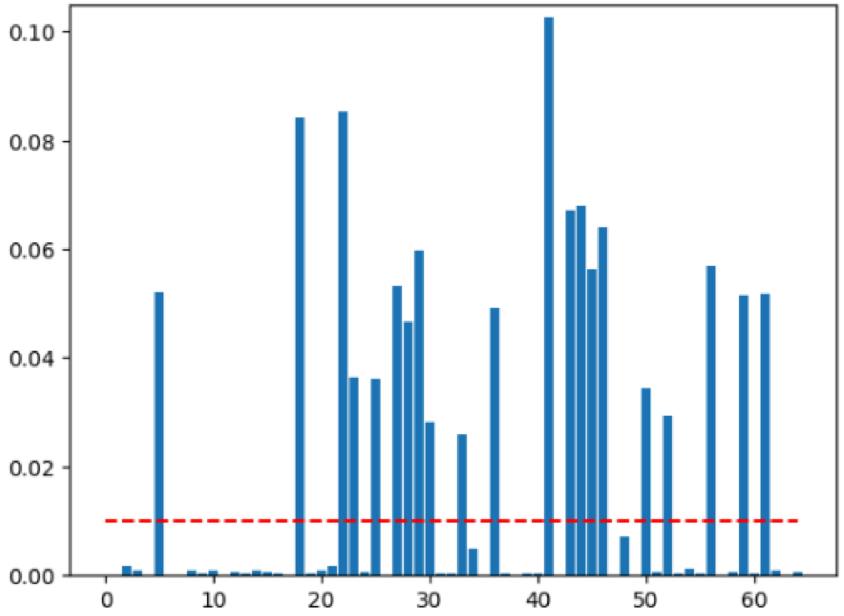
Empirical results



Soft gate



Gaussian prior

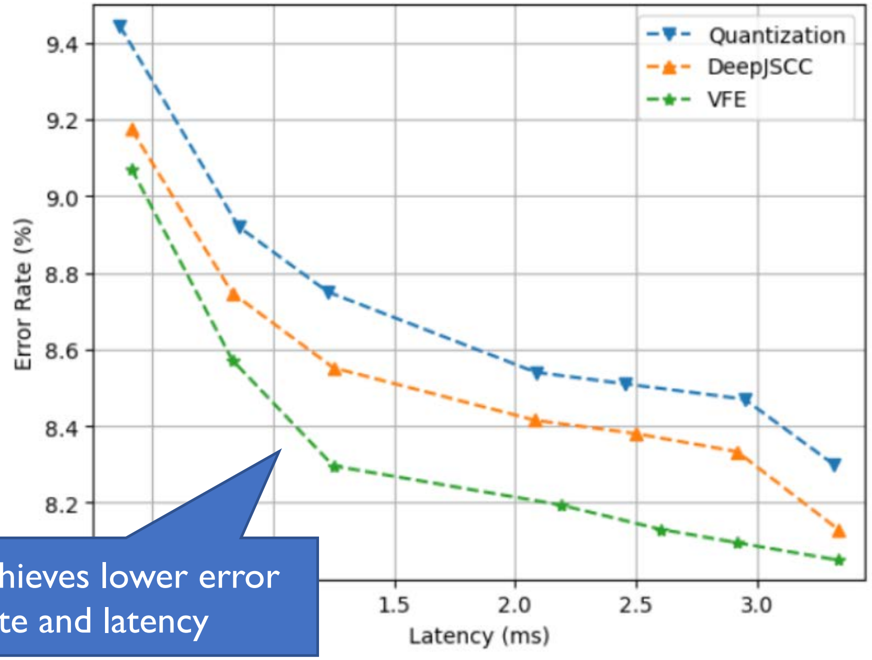


Log-uniform distribution

Experiment

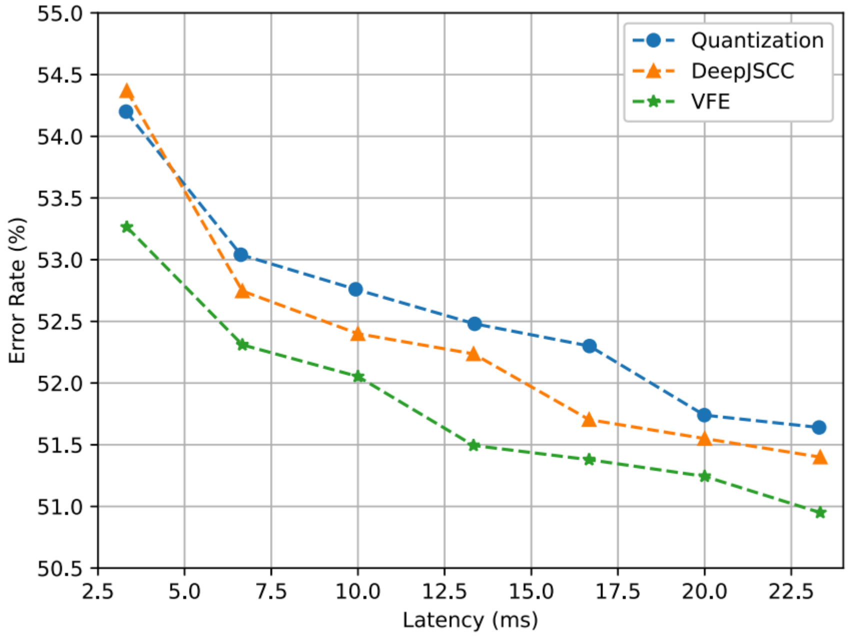
- **Baselines** (data-oriented communication):
  - DeepJSCC (Joint Source-Channel Coding)
  - Learning-based quantization (w/ ideal channel coding)

Rate-distortion on CIFAR-10 dataset



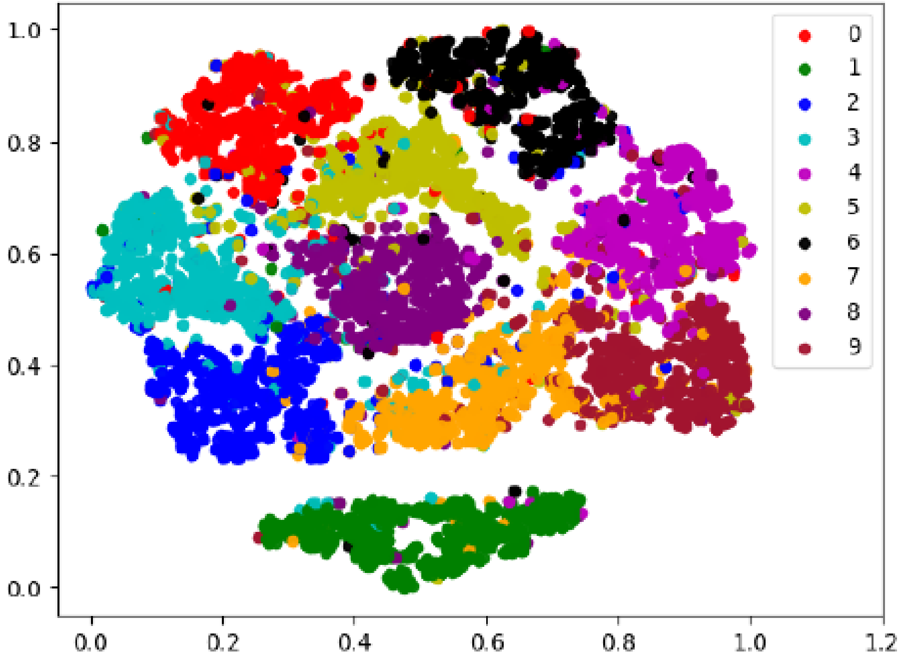
VFE achieves lower error rate and latency

Rate-distortion on Tiny ImageNet dataset

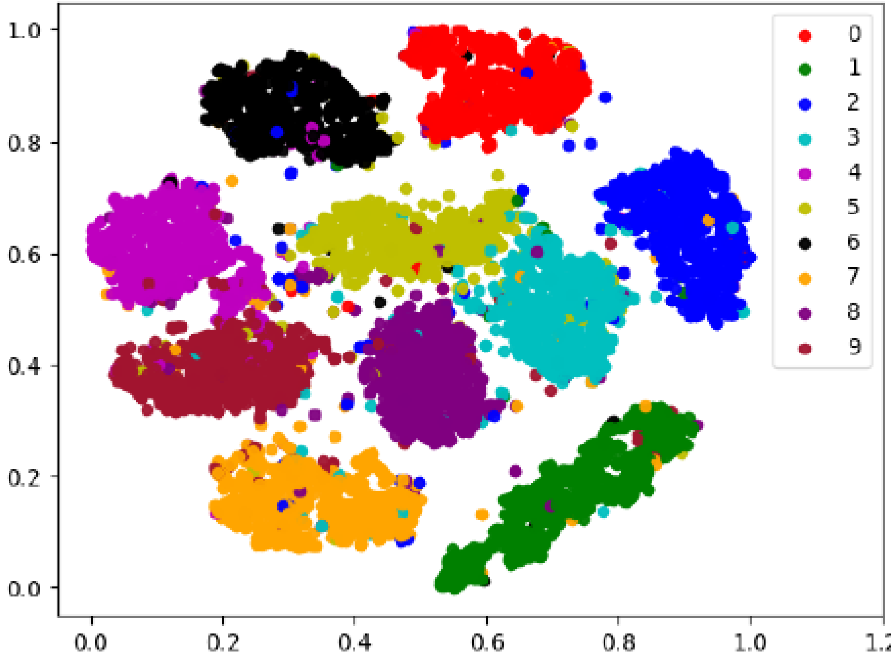


Experiment

- VFE method can better distinguish the data from different classes compared with DeepJSCC.



(a) DeepJSCC: Accuracy = 96.77%, dimension n = 24.

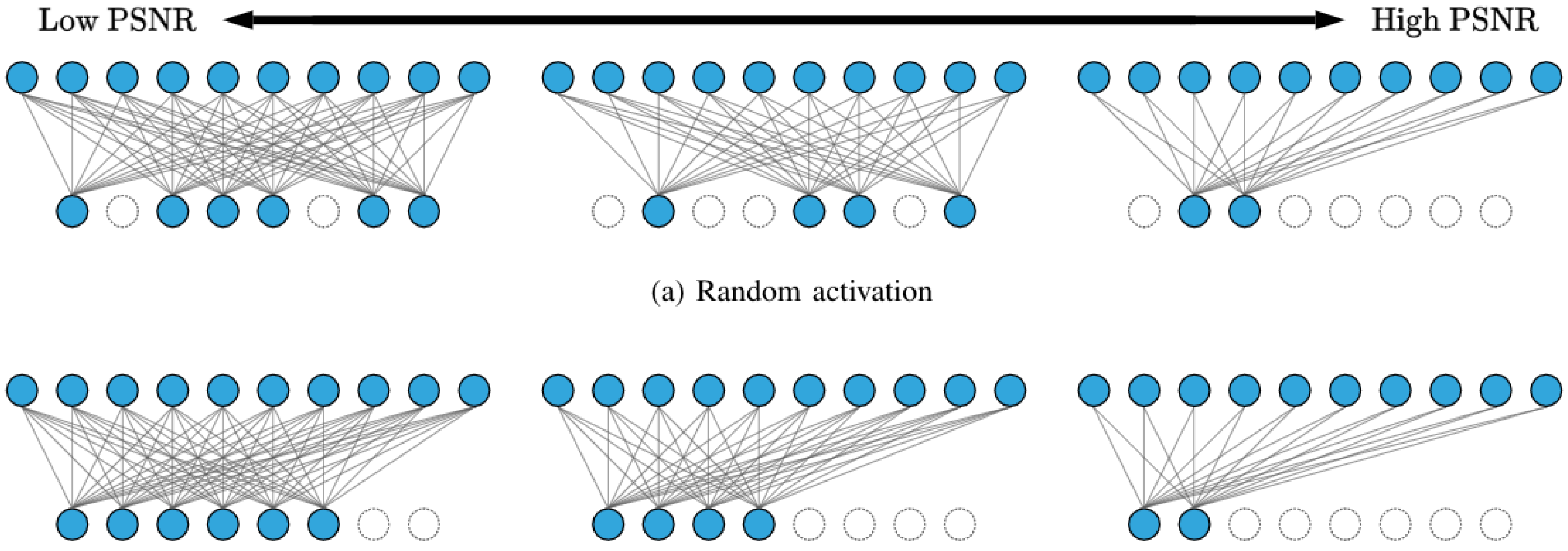


(b) Proposed VFE: Accuracy = 97.39%, dimension n = 24.

2-dimensional t-SNE embedding of the received feature in the MNIST classification task with PSNR = 20 dB.

Variable-length Variational Feature Encoding (VL-VFE)

- To adapt to channel states: variable-length coding
- To reduce signaling overhead, the coding scheme should be consecutive and monotonic



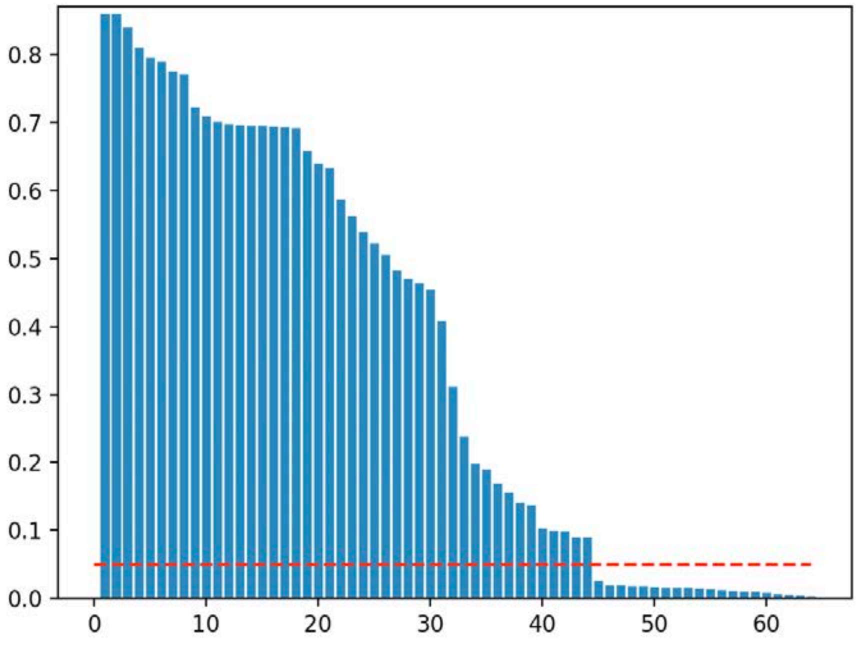
(a) Random activation

(b) Consecutive activation

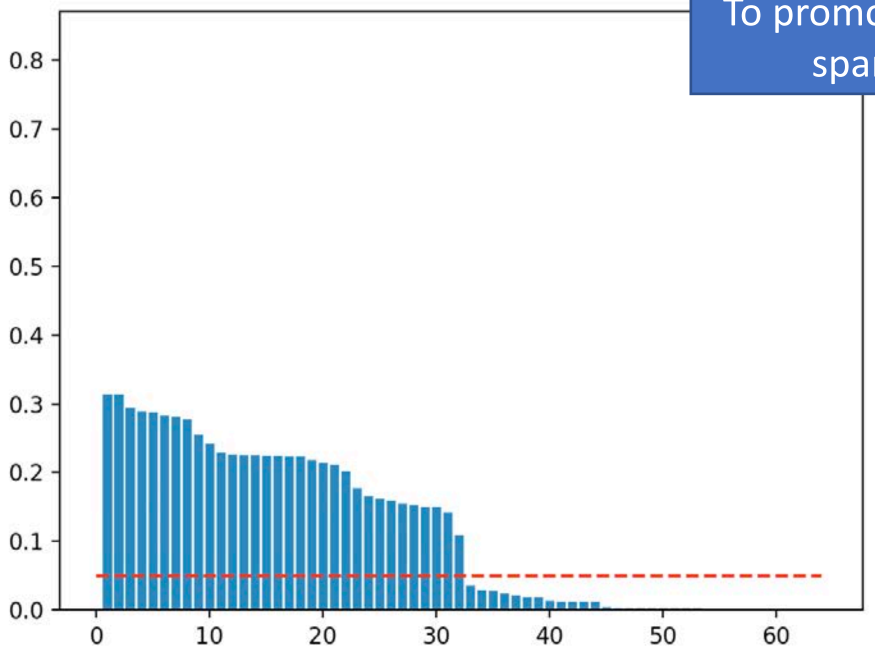
Variable-length Variational Feature Encoding (VL-VFE)

Dimension importance $\gamma_i(\sigma^2) = \sum_{j=i}^{n} |g_j(\sigma^2)|$ **Soft gate function**

To promote group sparsity

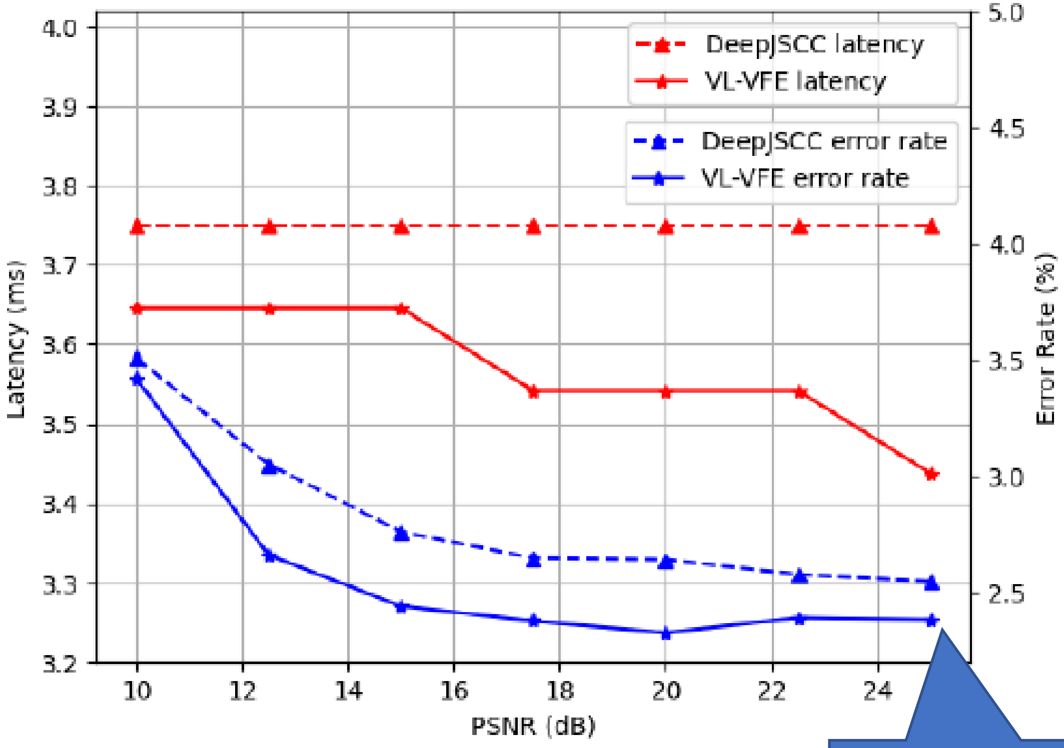


PSNR = 10dB

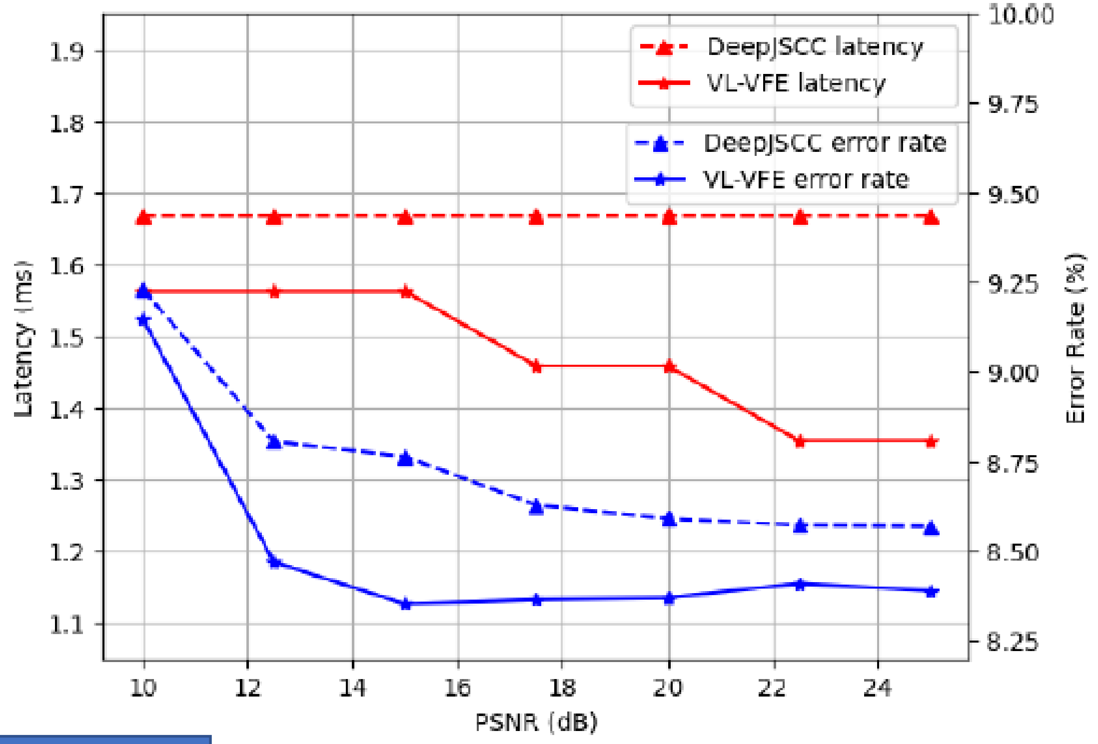


PSNR = 20dB

Variable-length Variational Feature Encoding (VL-VFE)



(a) The MNIST classification task



(b) The CIFAR-10 classification task

VL-VFE achieves lower error rate and latency

Task-oriented communication for cooperative inference via distributed information bottleneck

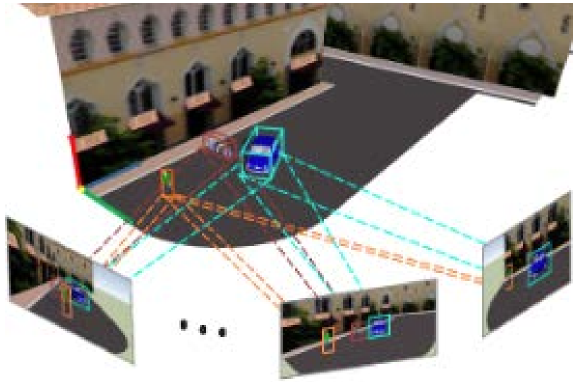
J. Shao, Y. Mao, and **J. Zhang**, "Task-oriented communication for multi-device cooperative edge inference," *IEEE Transactions on Wireless Communications*, vol. 11, no. 1, pp. 73-87, Jan. 2023.

New Applications: Cooperative Inference

- Multi-device system.
  - Cooperation among multiple devices with distinct views improves sensing capability.

Vehicle Re-identification

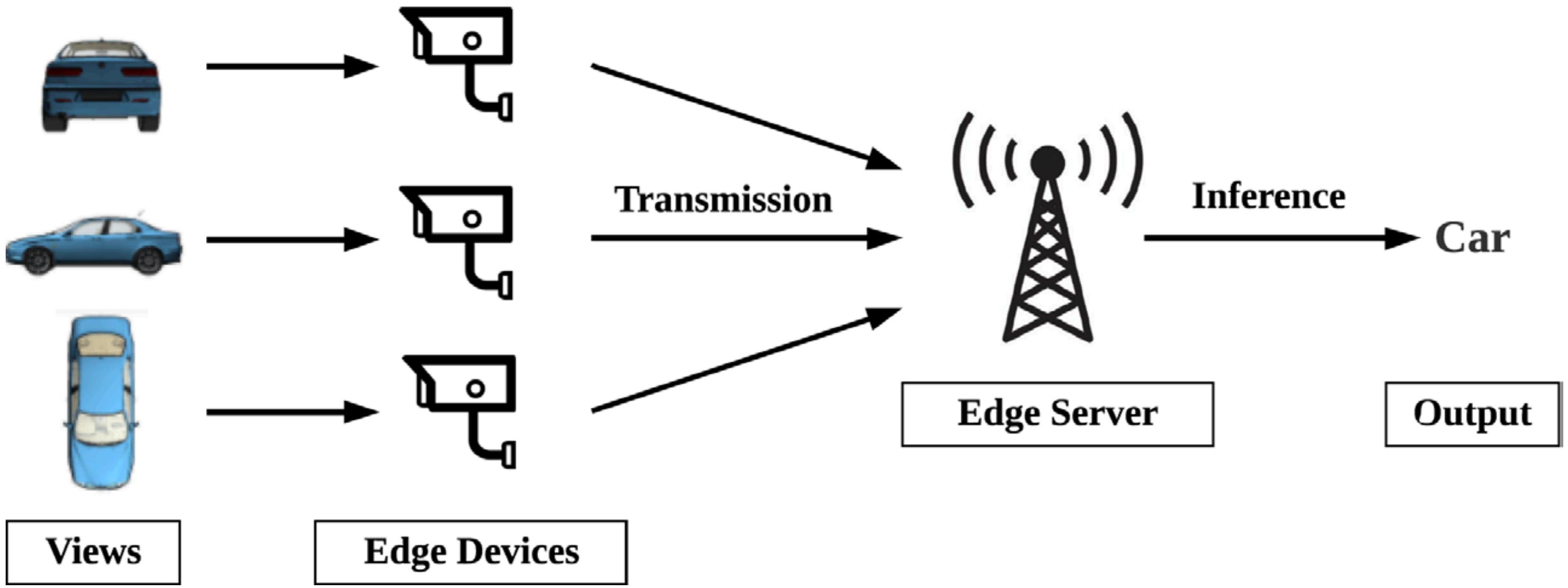
Pose Estimation



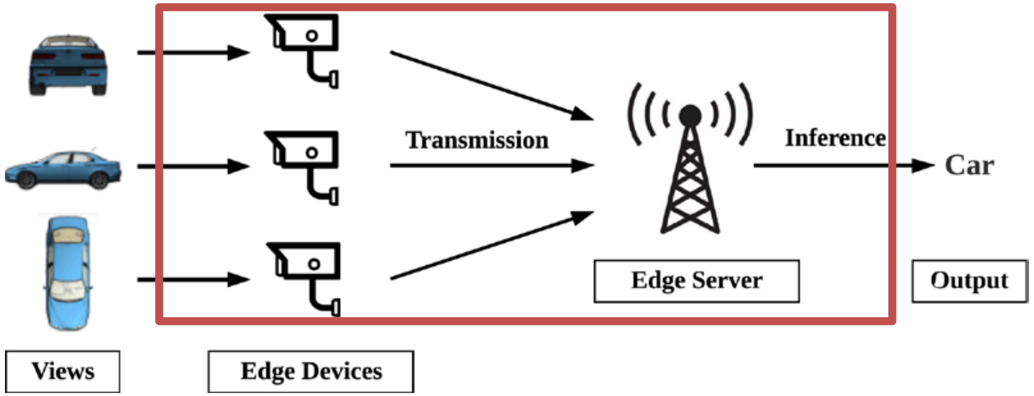
3D Localization

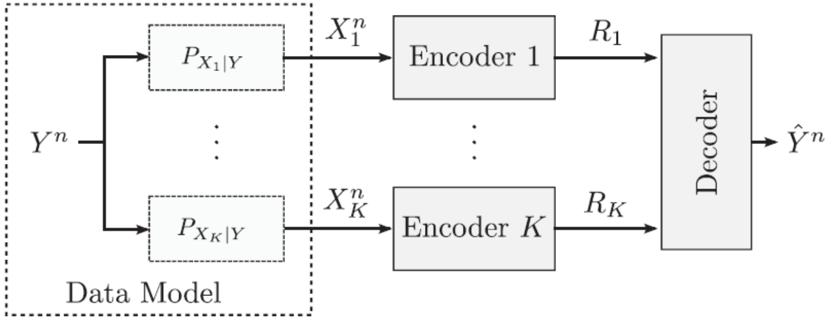
Multi-camera cooperative inference

- Objective: Design an efficient method that can fully exploit the correlation among multiple features in distributed feature encoding.



Cooperative perception vs. Distributed Information Bottleneck (DIB)





Distributed Information Bottleneck (DIB)

Closely related to the distributed Chief Executive Officer (CEO) source coding problem

Proposition. Suppose the input variables $X_k, \forall k = 1, 2, \ldots, K$ are conditional independet given Y. Given the relevance $\Delta = I(Y; Z_{1:K})$, the sum rate

Rate Relevance

$$\underbrace{\sum_{k=1}^{K} R_k}_{\text{Rate}} \geq \underbrace{\Delta}_{\text{Relevance}} + \sum_{k=1}^{K} [I(X_k; Z_k) - I(Y; Z_k)]$$

Relevance-rate tradeoff

Aguerri, Inaki Estella, and Abdellatif Zaidi. “Distributed variational representation learning.” *IEEE Trans. Pattern Anal. Machine Intell.* 120-138, 2019.

Multi-camera cooperative inference

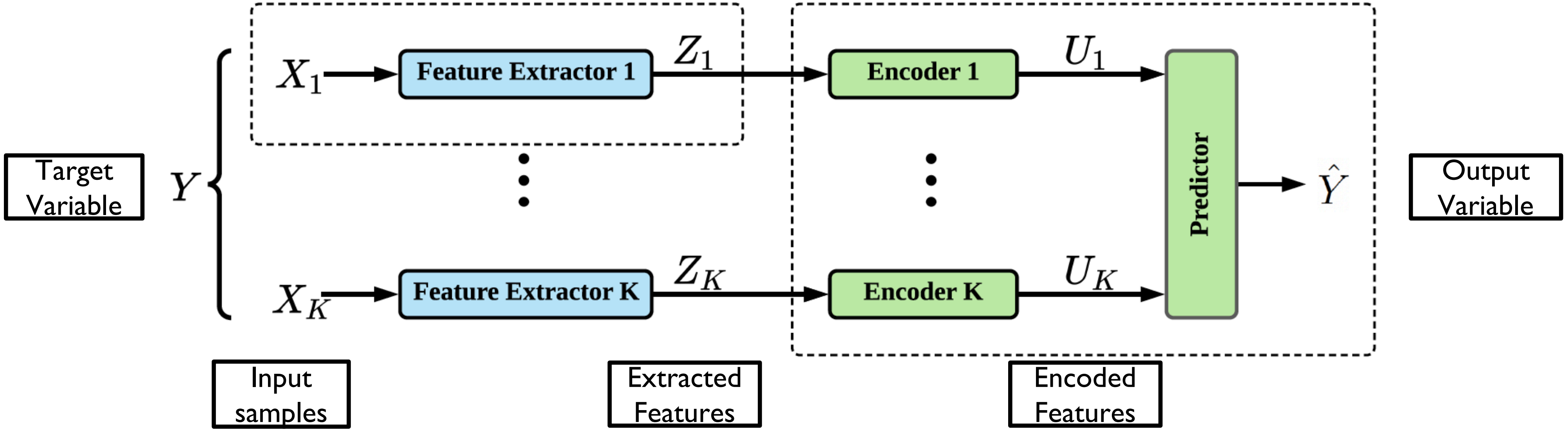
- Probabilistic modeling with *K* devices
- Loss functions

Information Bottleneck (IB)

$$\mathcal{L}_{\mathrm{IB,k}}(\gamma) := H(Y \mid Z_k) + \gamma I(X_k; Z_k)$$

Distributed IB (DIB)

$$\mathcal{L}_{\mathrm{DIB}}(\beta) := H(Y \mid U_{1:K}) + \beta \sum_{k=1}^{K} [H(Y \mid U_k) + I(Z_k; U_k)]$$



Distributed Deterministic Information Bottleneck (DDIB)

- DIB objective

$$\mathcal{L}_{\text{DIB}}(\beta) := H(Y \mid U_{1:K}) + \beta \sum_{k=1}^{K} [H(Y \mid U_k) + \underbrace{I(Z_k; U_k)}_{\text{Rate}}]$$

The minimality is only satisfied in the asymptotic limit

- DDIB objective

$$\mathcal{L}_{\text{DDIB}}(\beta) := H(Y \mid U_{1:K}) + \beta \sum_{k=1}^{K} [H(Y \mid U_k) + R_{\text{bit}}(U_k)]$$

Proposed method: Variational DDIB (VDDIB)

- Using variational inference to estimate the intractable (entropy) terms.

$$\mathcal{L}_{\mathrm{DDIB}}(\beta) := H(Y \mid U_{1:K}) + \beta \sum_{k=1}^{K} [H(Y \mid U_k) + R_{\mathrm{bit}}(U_k)]$$

$$\Downarrow$$

$$\mathcal{L}_{\mathrm{VDDIB}}(\beta; \boldsymbol{\phi}, \boldsymbol{\psi}) := \mathbf{E}_{p_\theta(\boldsymbol{z}_{1:K}, \boldsymbol{y})} \Big\{ -\log p_{\psi_0}(\boldsymbol{y} \mid \boldsymbol{u}_{1:K}) + \beta \Big\{ \sum_{k=1}^{K} -\log p_{\psi_k}(\boldsymbol{y} \mid \boldsymbol{u}_k) + \sum_{k=1}^{K} R_{\mathrm{bit}}(\boldsymbol{u}_k) \Big\} \Big\}$$

Minimizing the VDDIB objective may not result in the optimal rate-relevance tradeoff due to the approximations

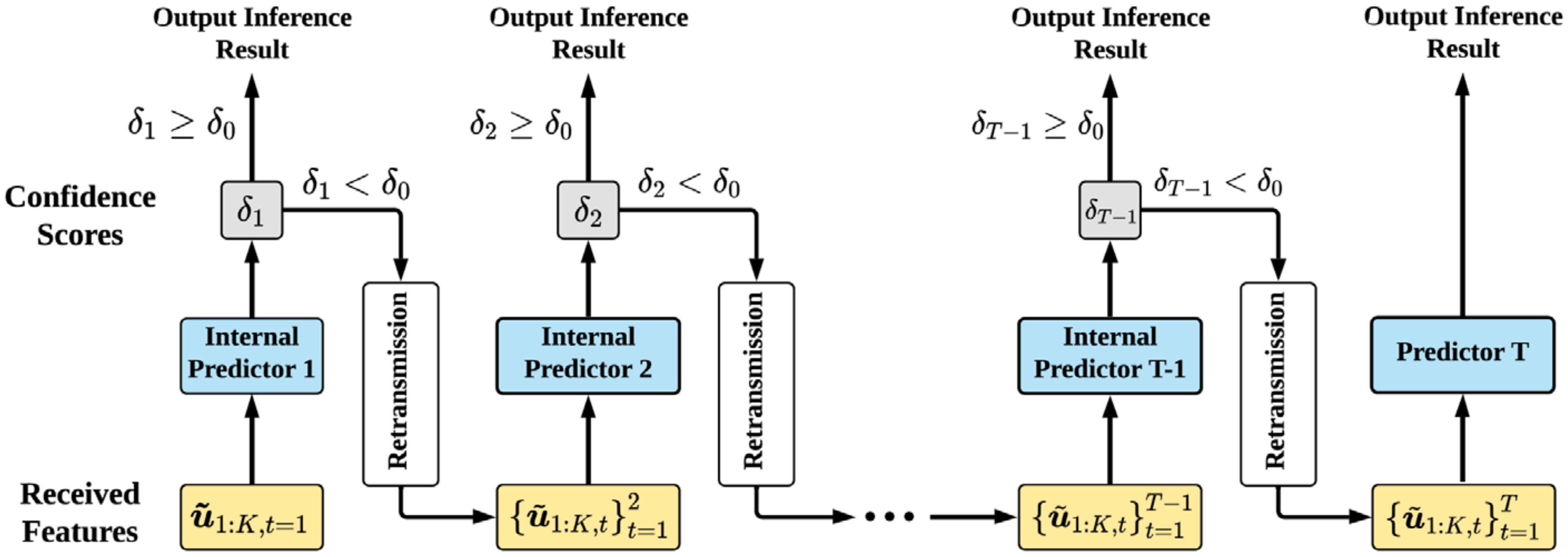
Introduce a selective retransmission (SR) mechanism to further reduce the communication overhead caused by the redundancy among the extracted features.

- The edge server selectively activates the edge devices to retransmit their encoded features based on the informativeness of the received features.
- The mechanism consists of a stopping policy and an attention module.

Selective Retransmission Mechanism

❖ Stopping policy

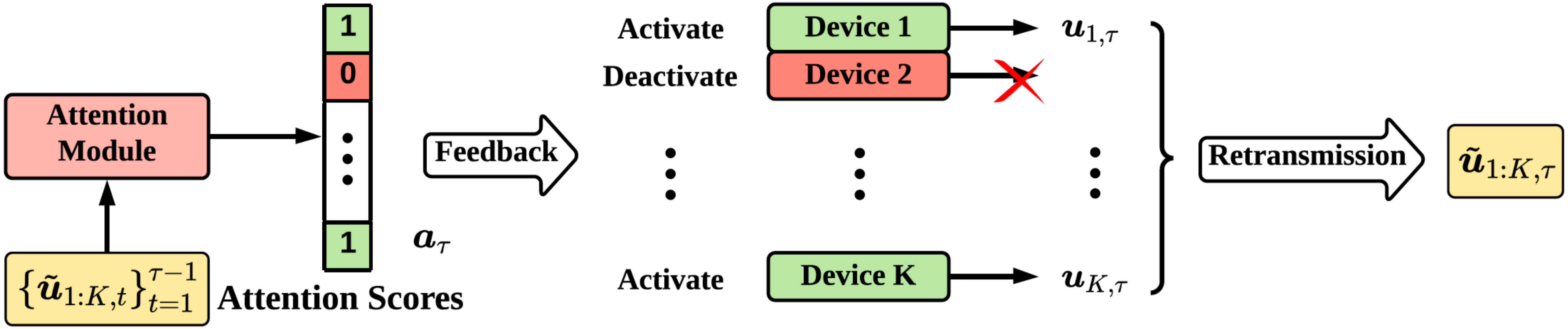
- Each edge device is allowed to transmit the encoded feature with a maximum number of T attempts.
- Once the received features are sufficient to output a confident result, the remaining retransmission attempts can be saved.



Selective Retransmission Mechanism

❖ Attention Module

- Select the most informative features to retransmit based on the attention scores.



VDDIB with Selective Retransmission Mechanism (VDDIB-SR)

- VDDIB-SR loss function

$$\mathcal{L}_{\mathrm{VDDIB}}(\beta;\boldsymbol{\phi},\boldsymbol{\psi}) := \mathbf{E}_{p_\theta(\boldsymbol{z}_{1:K},\boldsymbol{y})}\Big\{-\log p_{\psi_0}(\boldsymbol{y}\mid\boldsymbol{u}_{1:K}) + \beta\Big\{\sum_{k=1}^{K} -\log p_{\psi_k}(\boldsymbol{y}\mid\boldsymbol{u}_k) + \sum_{k=1}^{K} R_{\mathrm{bit}}(\boldsymbol{u}_k)\Big\}\Big\}$$

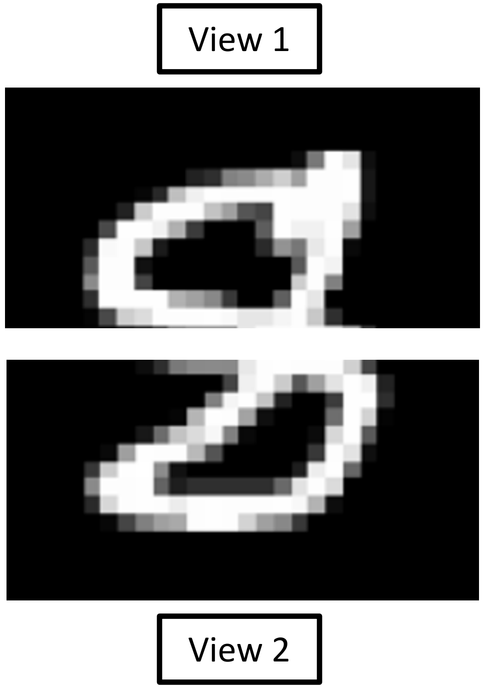
Account for T predictors

Communication cost by the SR mechanism

$$\mathcal{L}_{\mathrm{VDDIB-SR}}\left(\beta,T;\widetilde{\boldsymbol{\phi}},\widetilde{\boldsymbol{\psi}},\{\boldsymbol{\psi}_k\}_{k=1}^{K}\right) := \mathbf{E}_{p_\theta(\boldsymbol{z}_{1:K},\boldsymbol{y})}\left\{\frac{1}{T}\sum_{\tau=1}^{T} -\log p_{\tilde{\psi}_\tau}\left(\boldsymbol{y}\mid\{\tilde{\boldsymbol{u}}_{1:K,t}\}_{t=1}^{\tau}\right) + \beta\Big\{\sum_{k=1}^{K} -\log p_{\psi_k}(\boldsymbol{y}\mid\boldsymbol{u}_k) + \sum_{k=1}^{K}\sum_{t=1}^{T} R_{\mathrm{bit}}(\tilde{\boldsymbol{u}}_{k,t})\Big\}\right\}$$

Performance evaluation

- Cooperative inference tasks



Two-view MNIST classification

Twelve-view Shape Recognition on ModelNet40 dataset

Performance evaluation

- The accuracy of the cooperative tasks under different bit constraints.

- Task-oriented vs. Data-oriented
  - **MNIST classification task**: ~10 bits vs. 1.3 kbits
  - **Shape recognition task**: ~200 bits vs. 120 KB

MNIST classification

	R_{sum}		
	6 bits	10 bits	14 bits
NN-REG	95.93%	97.49%	97.78%
NN-GBI	96.62%	97.79%	98.02%
eSAFS	96.97%	97.87%	98.05%
CAFS	94.14%	97.43%	97.42%
VDDIB (ours)	97.08%	97.82%	98.06%
VDDIB-SR (T=2) (ours)	**97.13%**	**98.13%**	**98.22%**

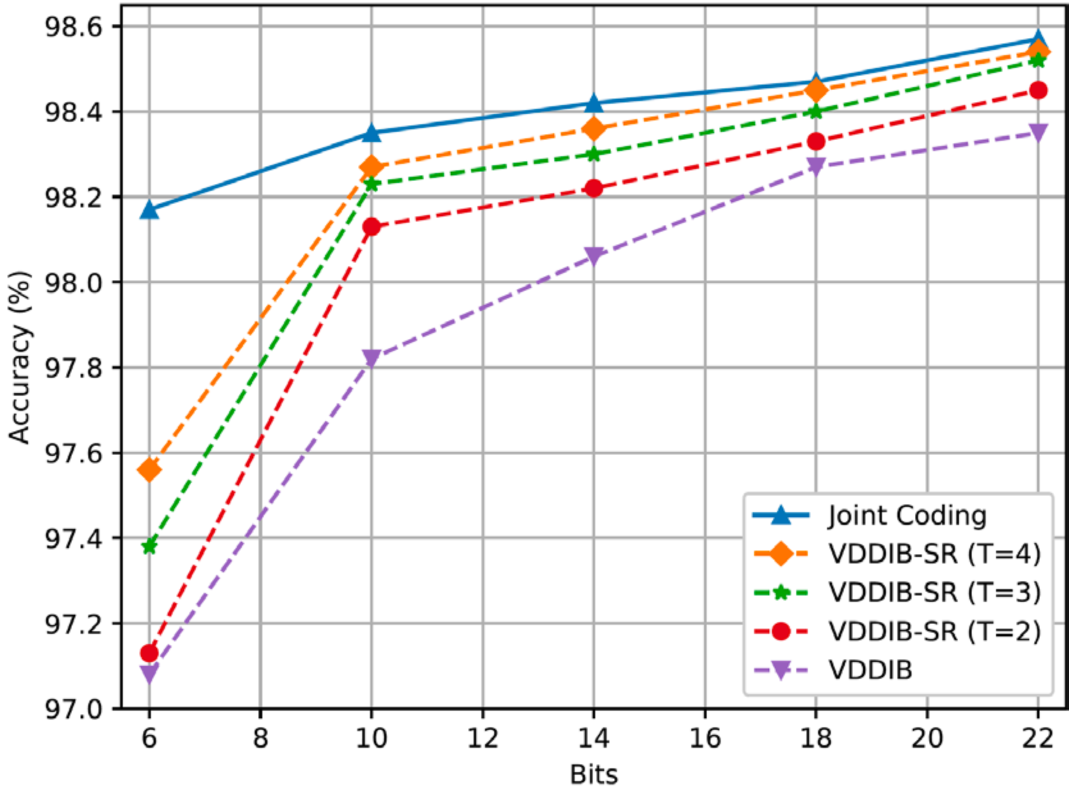
Shape Recognition

	R_{sum}		
	120 bits	240 bits	360 bits
NN-REG	87.50%	88.25%	89.03%
NN-GBI*	88.82%	—	—
eSAFS	85.88%	87.87%	89.50%
CAFS	86.75%	89.56%	90.67%
VDDIB (ours)	89.25%	90.03%	90.75%
VDDIB-SR (T=2) (ours)	**90.25%**	**91.31%**	**91.62%**

* The GBI quantization algorithm is computationally prohibitive when the number of bits is too large.

Ablation Study

- Impact of the maximum transmission attempts *T*.
  - the performance of the VDDIB-SR method improves *with T*.



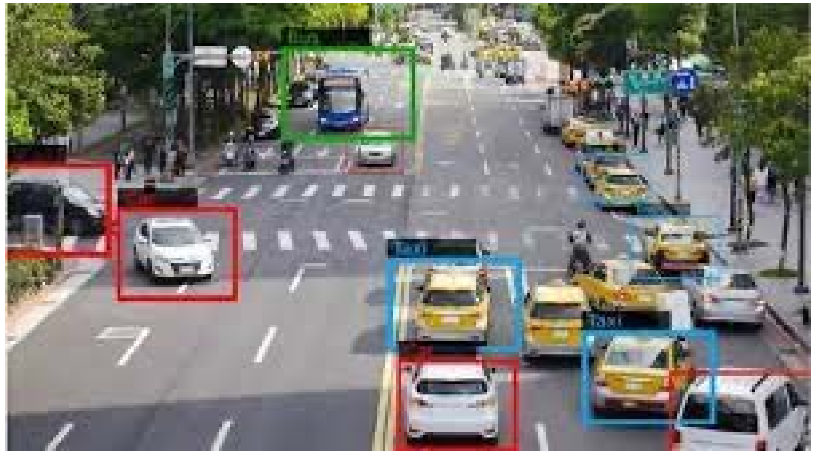
Task-oriented communication for edge video analytics (sequential data)

J. Shao, X. Zhang, and **J. Zhang**, "Task-oriented communication for edge video analytics," submitted to *IEEE Transactions on Wireless Communications*. (https://arxiv.org/abs/2211.14049)

Edge video analytics

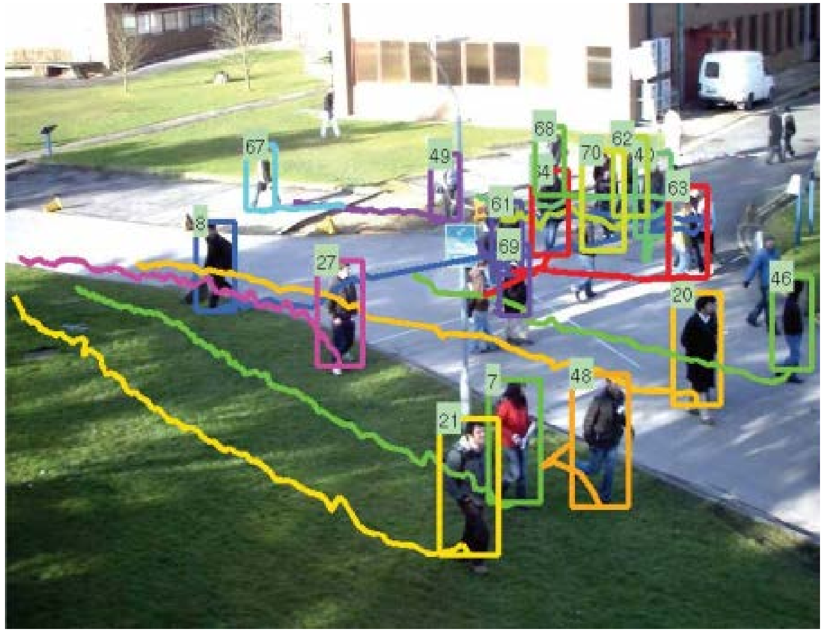
- More and more cameras and video data at the edge


- Powerful AI models for visual data

Sequence data processing at the network edge

- The observations are temporally correlated
  - Example: multi-camera surveillance system
  - Exploit the temporal correlation to reduce the communication overhead

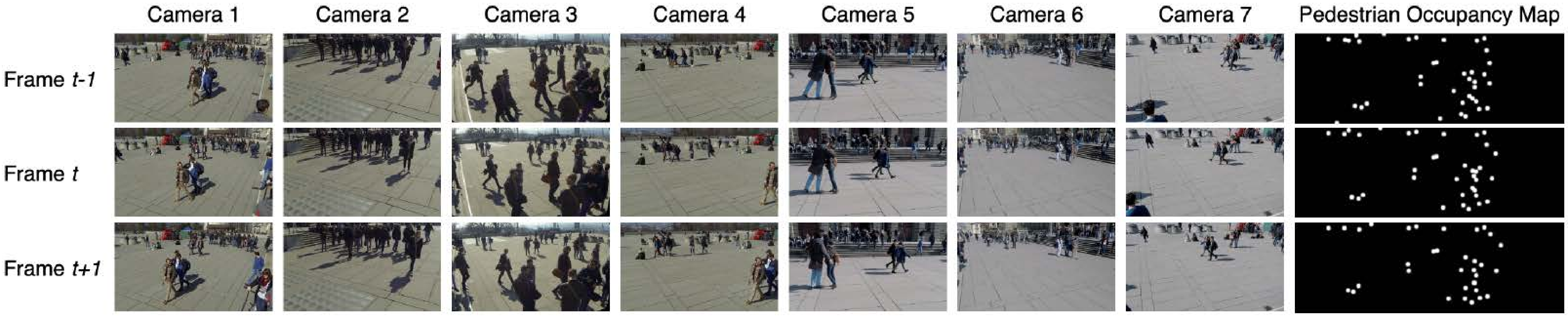


Single-camera tracking

Multi-target multi-camera re-identification

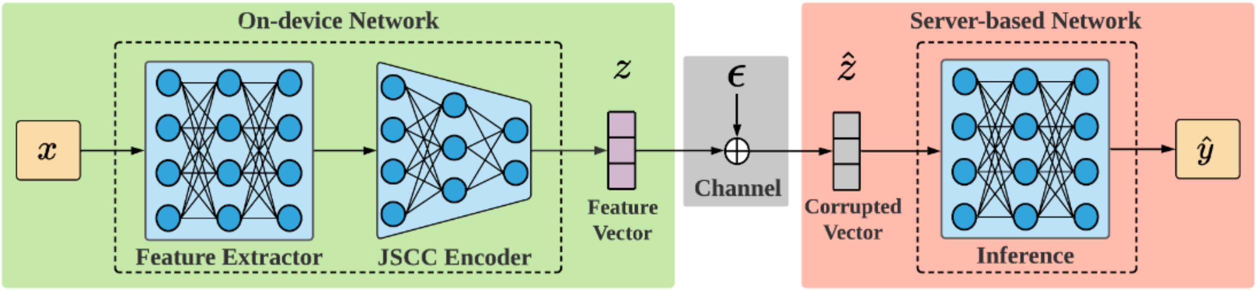
An example of edge video analytics

- Challenges in edge video analytics:
  - How to effectively exploit the **temporal dependence** among frames.
  - How to effectively leverage the **spatial correlation** among cameras.

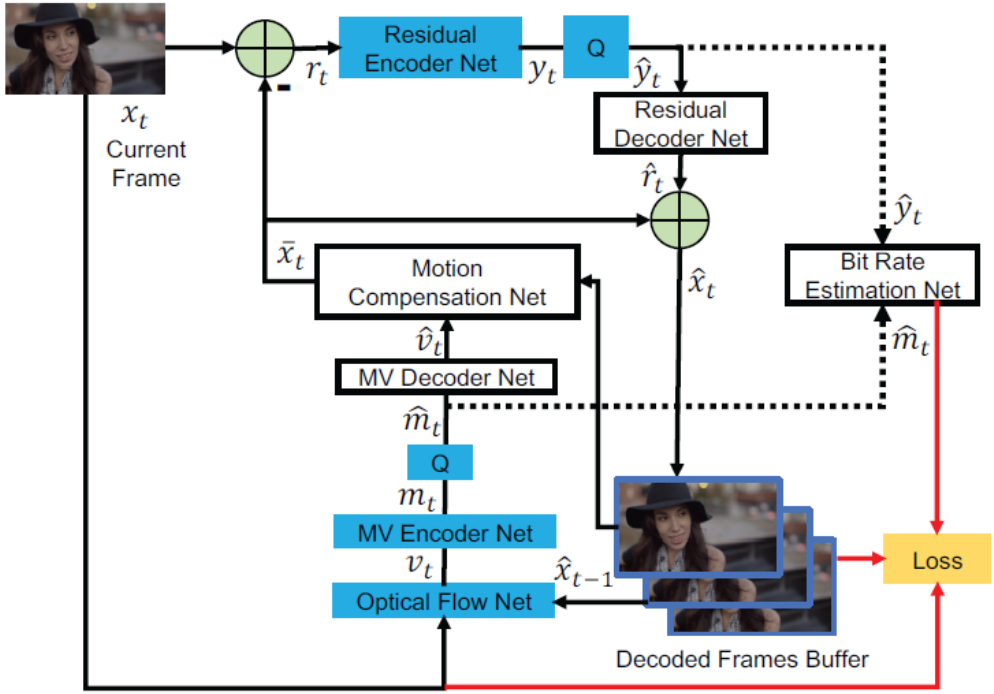


Existing methods

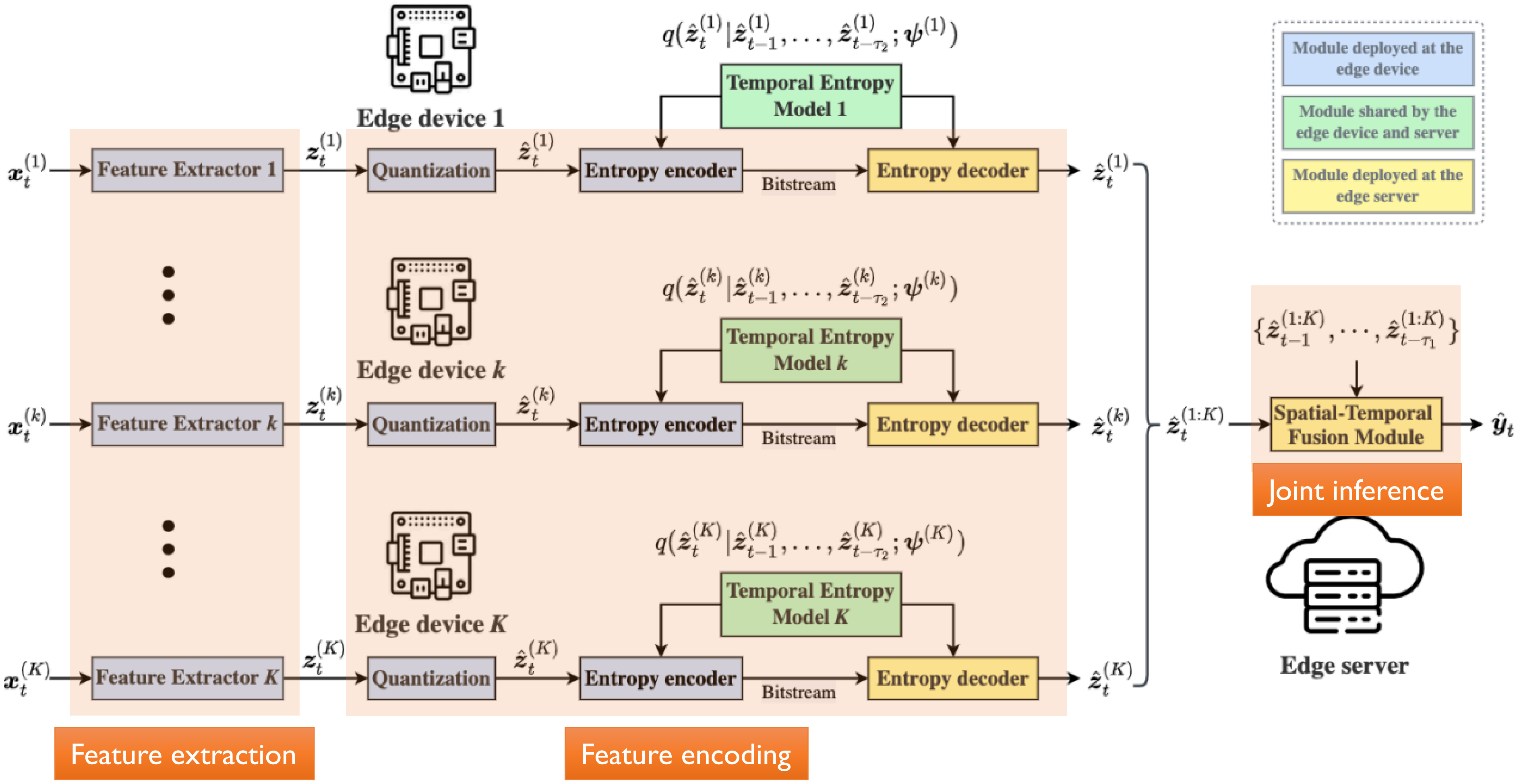
- **VFE**
  - Task-oriented
  - Only for images



- **DVC** (Deep video compression)
  - Efficient in extracting temporal correlation
  - But data-oriented



Proposed method



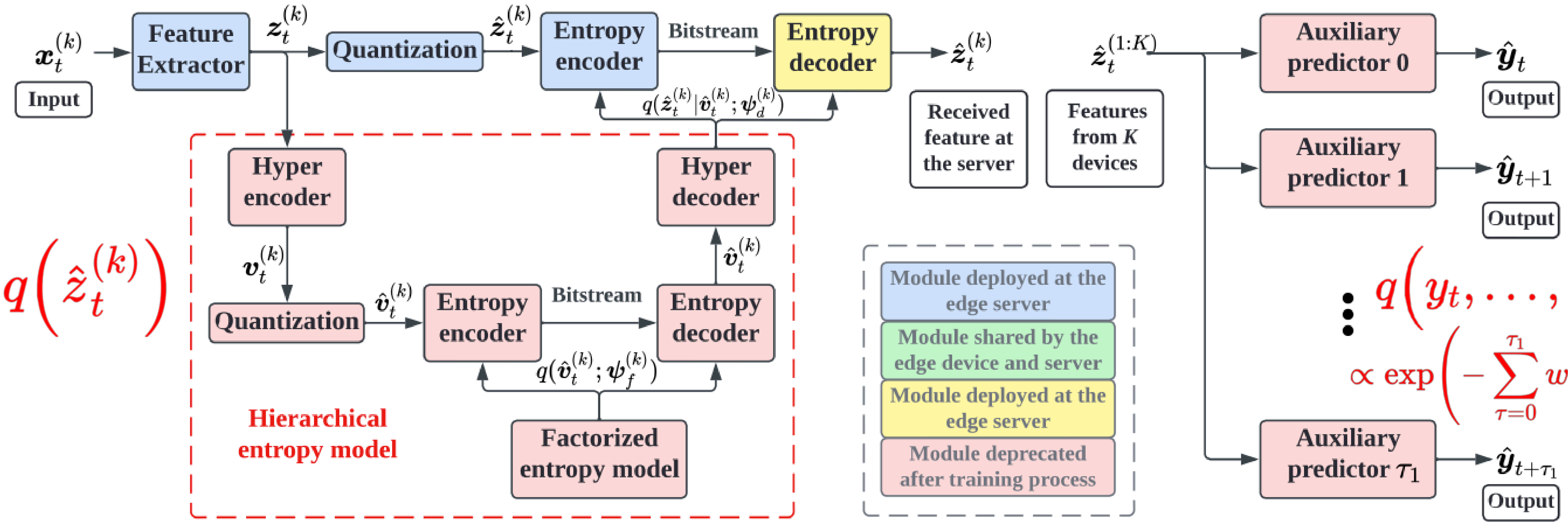
Feature extraction

$$X_t^{(k)} \to \hat{Z}_t^{(k)}$$

Input Feature

$$\min_{\boldsymbol{\theta}(1:K)} -\sum_{t=1}^{N} \boxed{I\left(Y_t,\ldots,Y_{t+\tau_1};\hat{Z}_t^{(1:K)}\right)} + \beta\sum_{t=1}^{N}\sum_{k=1}^{K} H\left(\hat{Z}_t^{(k)}\right)$$

$$p\left(y_t,\ldots,y_{t+\tau_1}\middle|\hat{z}_t^{(1:K)}\right) \qquad p\left(\hat{z}_t^{(k)}\right)$$



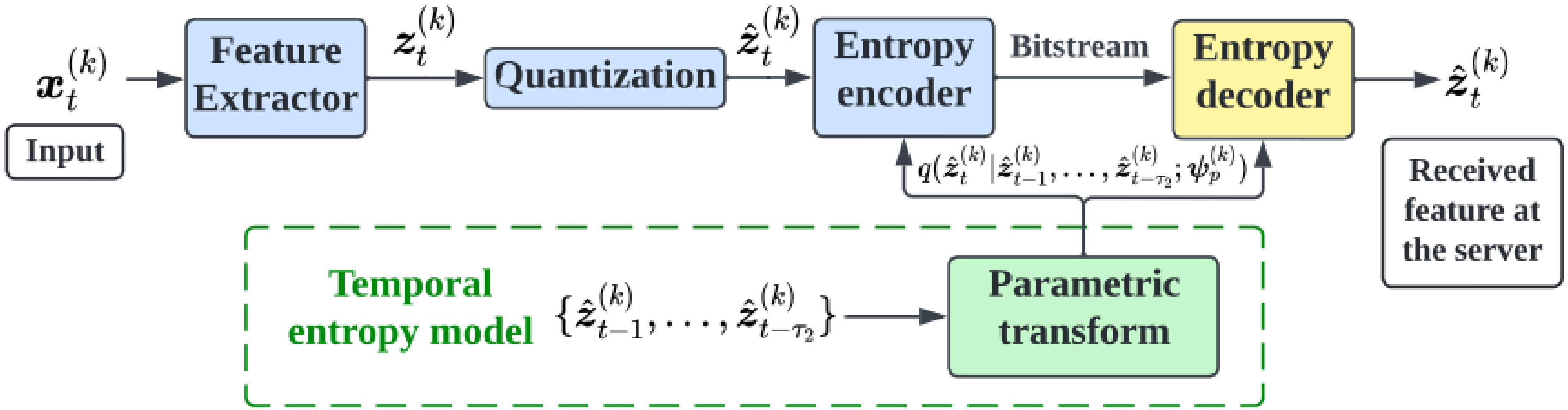
$$q\left(\hat{z}_t^{(k)}\right)$$

$$q\left(y_t,\ldots,y_{t+\tau_1}\middle|\hat{z}_t^{(1:K)}\right) \propto \exp\left(-\sum_{\tau=0}^{\tau_1} w_\tau d(\boldsymbol{y}_{t+\tau},\hat{\boldsymbol{y}}_{t+\tau};\boldsymbol{\varphi}_\tau)\right)$$

Feature encoding

- Temporal entropy model

$$H\left(\hat{\boldsymbol{z}}_t^{(k)} \mid \hat{\boldsymbol{z}}_{t-1}^{(k)}, \ldots, \hat{\boldsymbol{z}}_{t-\tau_1}^{(k)}\right).$$



$$H(p(\hat{\boldsymbol{z}}_t^{(k)}|\hat{\boldsymbol{z}}_{t-1}^{(k)}, \ldots, \hat{\boldsymbol{z}}_{t-\tau_1}^{(k)}), q(\hat{\boldsymbol{z}}_t^{(k)}|\hat{\boldsymbol{z}}_{t-1}^{(k)}, \ldots, \hat{\boldsymbol{z}}_{t-\tau_1}^{(k)}; \boldsymbol{\psi}_p^{(k)}))$$

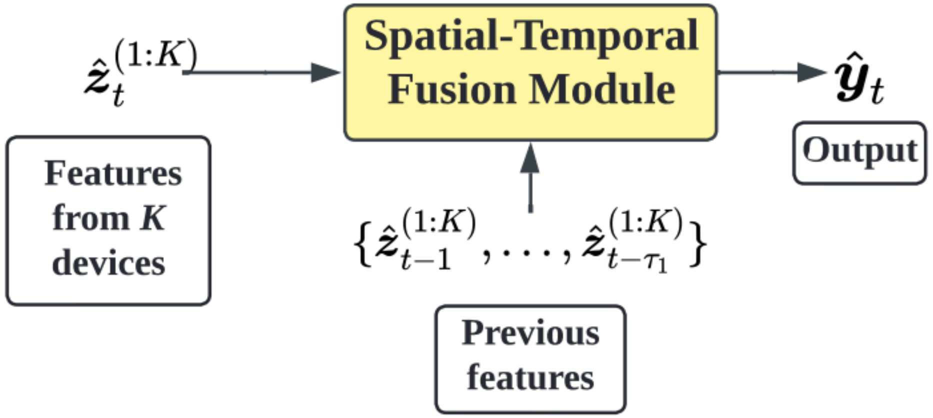
$$q(\hat{z}_{t,i}^{(k)}|\hat{\boldsymbol{z}}_{t-1}^{(k)}, \ldots, \hat{\boldsymbol{z}}_{t-\tau_1}^{(k)}; \boldsymbol{\psi}_p^{(k)}) = \left(\mathcal{N}(\mu_i^{(k)}, \sigma_i^{2(k)}) * \mathcal{U}(-0.5, 0.5)\right)(\hat{z}_{t,i}^{(k)})$$

$$\text{with } \boldsymbol{\mu}^{(k)}, \boldsymbol{\sigma}^{(k)} = h_p(\hat{\boldsymbol{z}}_{t-1}^{(k)}, \ldots, \hat{\boldsymbol{z}}_{t-\tau_1}^{(k)}; \boldsymbol{\psi}_p^{(k)}).$$

Joint inference (spatial-temporal fusion)

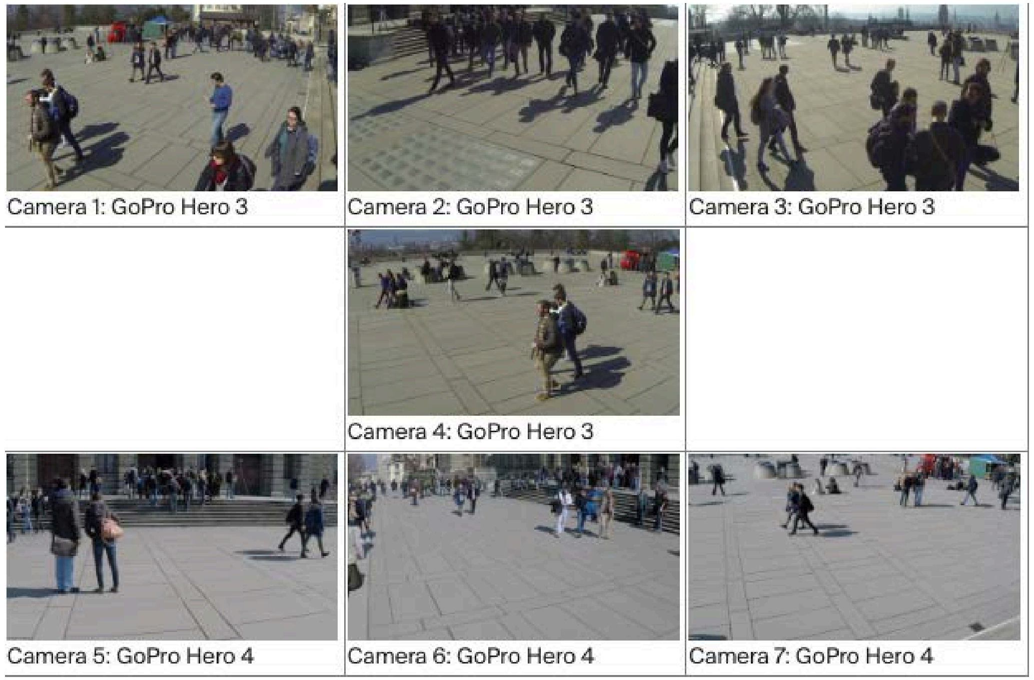
- Joint prediction module

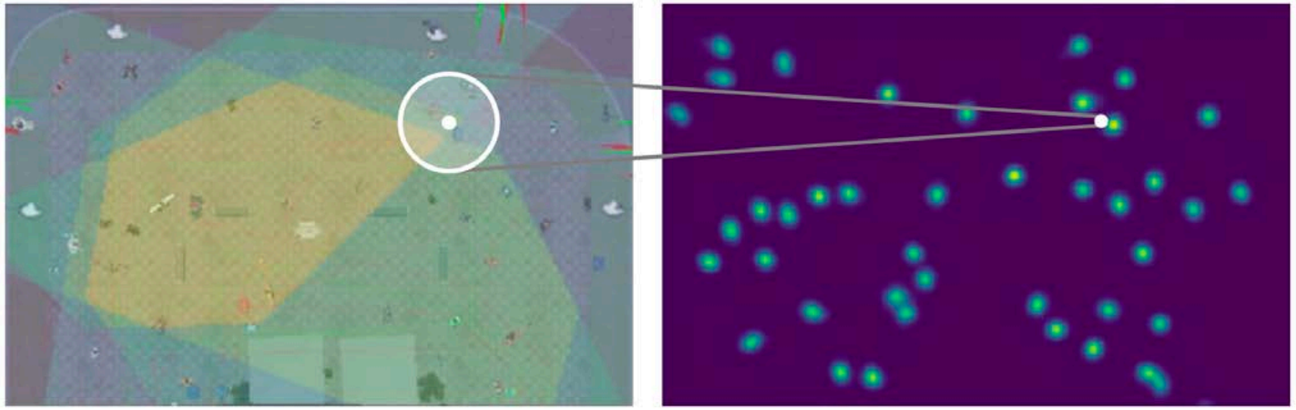
$$\hat{\boldsymbol{y}}_t = \boldsymbol{g}\left(\hat{\boldsymbol{z}}_t^{(1:K)}, \ldots, \hat{\boldsymbol{z}}_{t-\tau_1}^{(1:K)}; \boldsymbol{\phi}\right)$$



Experimental results

- Multi-camera pedestrian occupancy prediction (Wildtrack dataset)

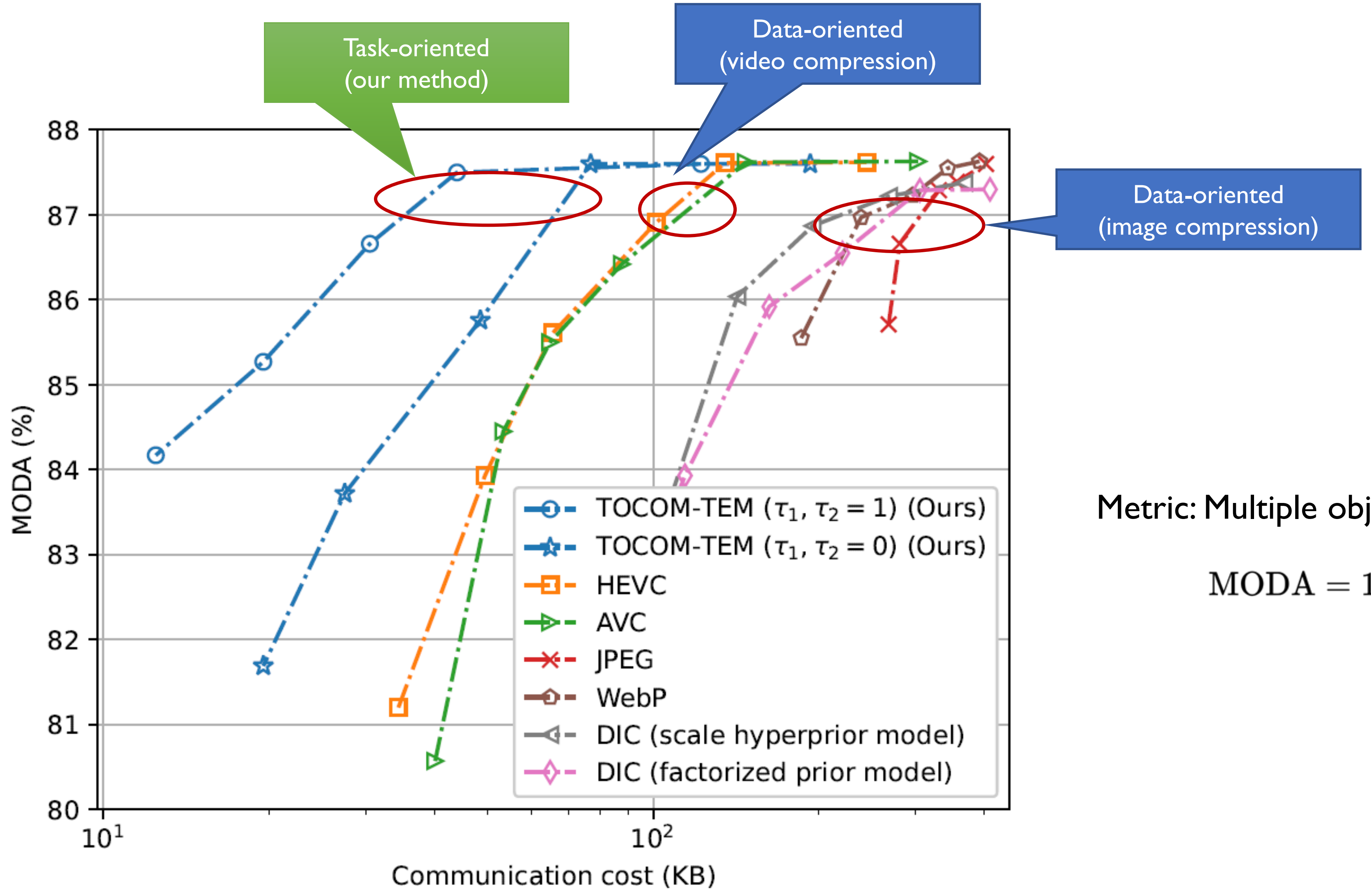




ground plane (bird’s eye view) Output: pedestrian occupancy map

- Chavdarova, Tatjana, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. "Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5030-5039. 2018.

Multi-camera pedestrian occupancy prediction

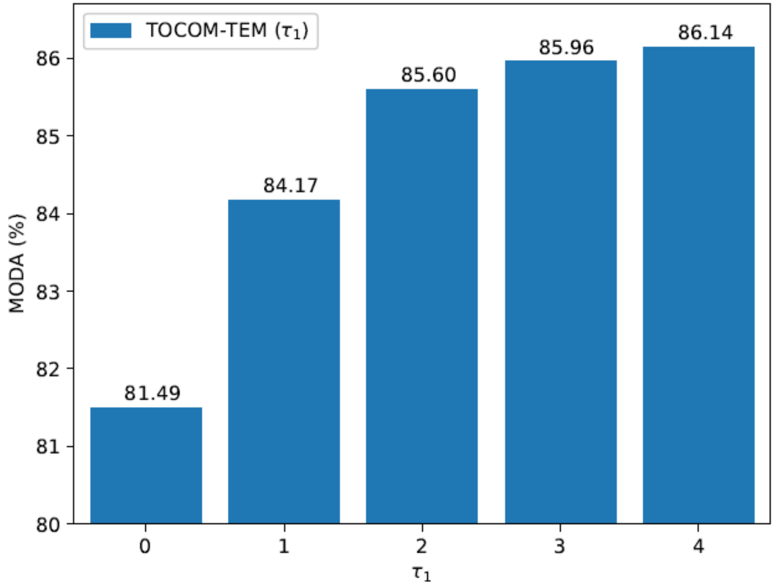


Metric: Multiple object detection accuracy

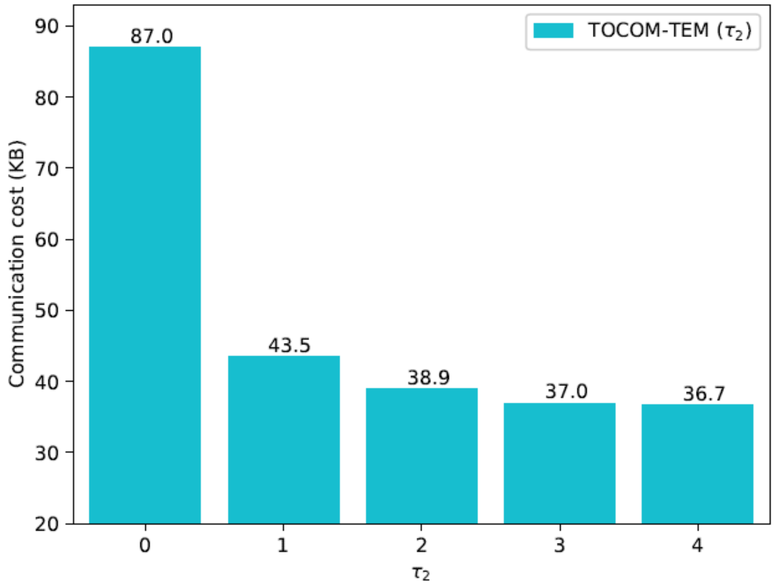
$$\mathrm{MODA} = 1 - \frac{FN + FP}{FN + TP}$$

Ablation study

Impact of the value of the parameter τ_1



Impact of the value of the parameter τ_2

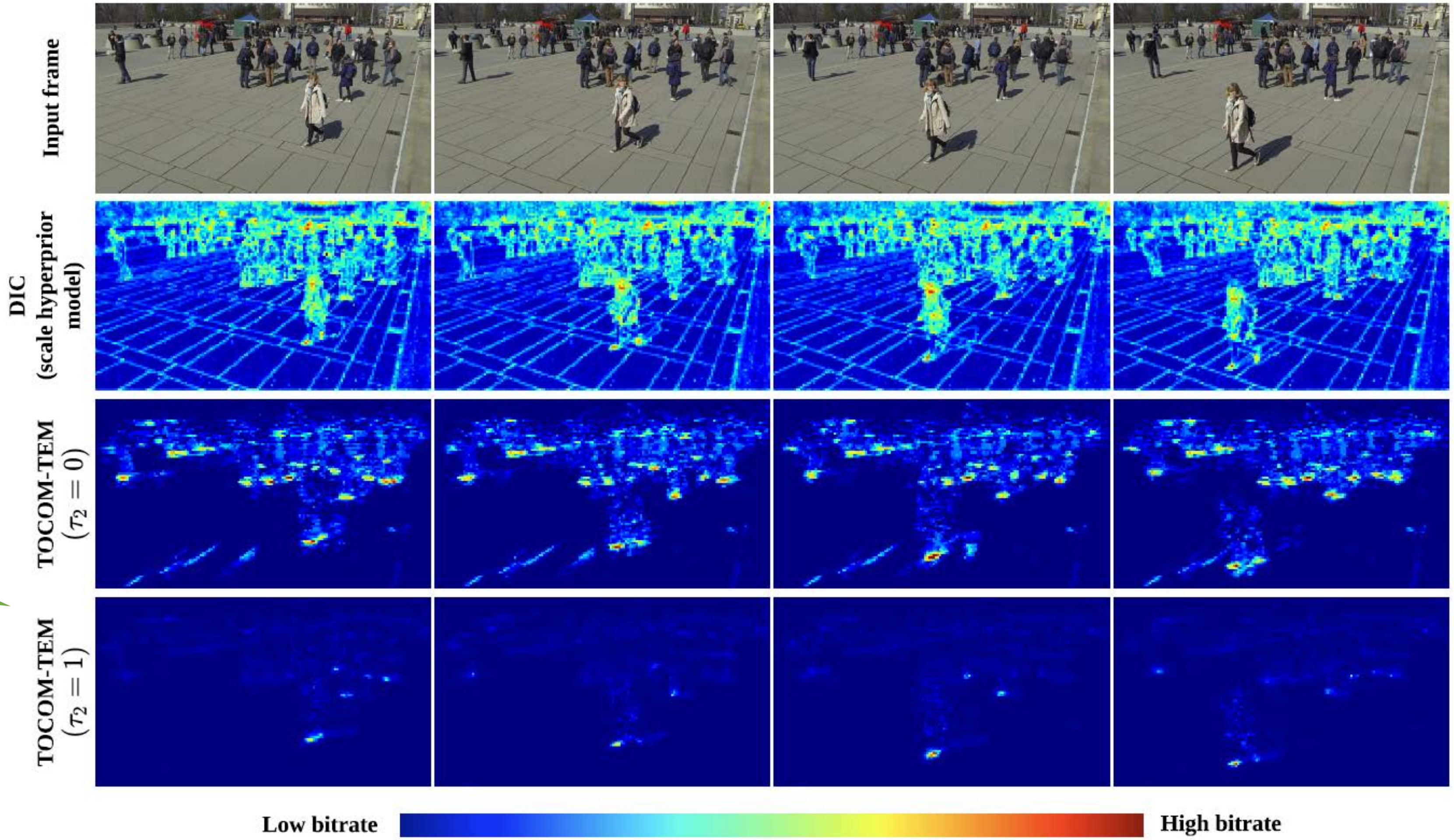


Ablation study

- Bit allocation of the transmitted features of different methods for the multi-camera pedestrian detection task.

Data-oriented (image compression)

Task-oriented (our method)



Conclusions

Conclusions

- Task-oriented communication
  - Shift from “how to communicate” to “what to communicate”
- Task-oriented communication for edge video analytics
  - Edge-assisted inference via information bottleneck
  - Cooperative perception via distributed information bottleneck
- Powerful tools
  - End-to-end optimization
  - Variational approximation

References

- J. Shao, **J. Zhang**, "BottleNet++: An end-to-end approach for feature compression in device-edge co-inference systems," *IEEE Int. Conf. Commun. (ICC) Workshop* 2020, June 2020.
- J. Shao, **J. Zhang**, "Communication-computation trade-off in resource-constrained edge inference," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 20–26, Dec. 2020.
- J. Shao, H. Zhang, Y. Mao, and **J. Zhang**, "Branchy-GNN: a device-edge co-inference framework for efficient point cloud processing," in Proc. *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Toronto, Ontario, Canada, Jun. 2021.
- J. Shao, Y. Mao, **J. Zhang**, "Learning task-oriented communication for edge inference: An information bottleneck approach," *IEEE J. Select. Areas Commun.*, vol. 40, no. 1, pp. 197-211, Jan. 2022.
- J. Shao, Y. Mao, and **J. Zhang**, "Task-oriented communication for multi-device cooperative edge inference," *IEEE Trans. Wireless Communications*, vol. 11, no. 1, pp. 73-87, Jan. 2023.
- J. Shao, X. Zhang, and **J. Zhang**, "Task-oriented communication for edge video analytics," submitted to *IEEE Transactions on Wireless Communications*. (https://arxiv.org/abs/2211.14049)

Thank you!

- For more details

  https://eejzhang.people.ust.hk/